NONPARAMETRIC AND SEMIPARAMETRIC METHODS FOR ECONOMIC RESEARCH

Miguel A. Delgado
Universidad Carlos III
and
Peter M. Robinson
London School of Economics

Abstract. Developments in the vast and growing literatures on nonparametric and semiparametric statistical estimation are reviewed. The emphasis is on useful methodology - rather than statistical properties for their own sake. Some empirical applications to economic data are described. The paper deals separately with nonparametric density estimation, nonparametric regression estimation, and estimation of semiparametric models.

Keywords. Nonparametric density estimation; nonparametric regression estimation; semiparametric models

1. Introduction

Econometrics is concerned with drawing statistical inferences from economic data. Statistical inferences must be based on a probability model for the data. The probability model may describe the joint distribution of the data, or it may only describe the conditional distribution of one set of observables given values of another set, or some aspect of this conditional distribution such as the conditional expectation (regression function).

In econometrics the probability model has most usually been parametric, that is, a given function involving a finite number of unknown parameters. In particular, a linear parametric function is often assumed. If the number of parameters is small relative to the number of observations, precise estimation of the parameters is possible, and consequently, reliable statistical inferences. Because many economic data sets are usually small, for example annual postwar macroeconomic time series, a parametric approach is sometimes essential. However, it is important that the parametric model be accurately chosen. Consistent parameter estimation generally requires an exactly correct choice of parametric model. Of course, this is never possible in practice. But economic theory may be insufficient to provide much confidence in any given parametric model as even a good approximation. Given a candidate set of explanatory

variables, a purely arbitrary functional form for a regression function is often used in much applied research, usually a form which is linear in the parameters owing to its desirable computational implications.

A nonparametric model makes no precise assumptions about functional form. Instead, the data are allowed to 'speak for themselves'. A nonparametric model provides a more robust approach to statistical inference because it is more likely to approximately capture the true underlying structure. With a very large amount of data, good estimates of a nonparametric model can be obtained. A nonparametric approach is especially useful at an exploratory level, to provide rough indication of which variables are relevant to the analysis of a particular problem, or of functional form of a regression model, or of distributional form of a disturbance random variable. Nowadays many large economic data sets are available, such as cross-sectional survey data consisting of thousands of observations, or intra-daily financial time series recorded at fine intervals of time, and these provide scope for reasonably precise estimation of a number of nonparametric models.

Semiparametric models provide a compromise between parametric and nonparametric models. A semiparametric probability model has two components, a parametric and a nonparametric one. Interest usually focuses on the estimation of the parametric component, the generality afforded by the nonparametric component providing a more robust environment for this than a pure parametric model. Because parameters, such as regression coefficients, may have a ready economic interpretation, there is some advantage over the nonparametric approach in retaining some element of finite parameterization. Semiparametric estimation is likely to require more data than parametric estimation. On the other hand it is likely to require less data than nonparametric estimation, indeed a satisfactorily precise nonparametric analysis involving a large number of explanatory variables would likely require an astronomically large sample, larger than any likely to be available in economic problems, and larger than any that we might have the resources to process.

This paper attempts to survey useful developments in nonparametric and semiparametric estimation. The literature is vast and rapidly growing, and so a comprehensive bibliography, let alone a full account of this literature, would be out of the question. In section 2 we discuss estimation of the probability density function. While of some direct interest in itself in economic research, our discussion of this topic also introduces themes relevant to nonparametric regression analysis, which is discussed in section 3, and both nonparametric density and regression estimates feature in semiparametric estimation, which is discussed in section 4. The paper places some stress on one important topic on which much of the progress has been recent, namely data-driven choice of smoothing numbers. We also refer to some economic applications of the various methods. Published empirical applications to economic data are not widespread, but the paper is written with the expectation that as the methodology becomes more widely known and better understood, it will find greater use by applied economists.

2. Density estimation

Probability density estimates are useful in exploratory data analysis of econometric data sets. A number of statistical procedures use density and derivative-of-density estimates; e.g. discriminant and cluster analysis; the estimation of probabilities, hazard rates, conditional densities and score functions; simulation; testing for unimodality and independence etc. Our discussion of this topic will also involve themes useful in sections 3 and 4 below.

The density estimation problem consists of estimating the functional form of the density from data. That is, the density f(.) of a r-valued random vector X is estimated by $\hat{f}_n(.)$ from data $\{X_1, X_2, ..., X_n\}$. Unless otherwise stated the X_i are assumed independent in what follows.

One approach to density estimation is *parametric*. Assuming f belongs to a parametric family of densities, the parameters can be estimated, for example by maximum likelihood, from the observed data set. Economic models usually do not justify a precise parameterization but they may provide information on certain features of the density shape; e.g. skewness, kurtosis, multimodality, monotonicity, etc. For instance, it is known, from casual observation, that income distributions are skewed to the right like the lognormal and gamma distributions and a vast number of mixture distributions.

Nonparametric estimation provides a way of avoiding the imposition of a rigid functional form on the density *a priori*. Hildenbrand and Hildenbrand (1982) compared nonparametric estimates of the income distribution with maximum likelihood based on the lognormal and gamma densities. They found that nonparametric density estimates are indeed skewed to the right but that their shape is very different from those of the lognormal and gamma. The literature on nonparametric density estimation is immense. Some books on the topic are Tapia and Thompson (1978), Hand (1982), Prakasa Rao (1983), Devroye and Györfi (1985), Silverman (1986) and Devroye (1987); some survey papers are Rosenblatt (1971), Wegman (1972), Tarter and Kronmal (1976), Fryer (1977), Leonard (1978), Bean and Tsokos (1980) and Izenman (1991); and an extensive (but inevitably out-of-date) bibliography is in Wertz and Schneider (1979). We next present some density estimation techniques used in nonparametric estimation and then we discuss some applications in economics.

2.1. *Some techniques*

The traditionally most popular method is the histogram. In order to construct a histogram, the data set $\{X_1, X_2, ..., X_n\}$ is divided into a number, n, of partitions $A_{n1}, A_{n2}, ..., A_{nm}$ and $f(x)$ is estimated by

$$\hat{f}(x) = n^{-1} \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{1(X_j \in A_{ni}) 1(x \in A_{ni})}{\lambda(A_{ni})}, \tag{1}$$

where 1(.) is the indicator function and $\lambda(A)$ is the Lebesgue measure of the partition A, i.e. the length of the interval when $r = 1$, the area when $r = 2$, the

volume when r = 3 or the hypervolume when r > 3. When n is large, (1) is a good approximation to

$$\sum_{i=1}^{m} \Pr\{X \in A_{ni} \text{ and } x \in A_{ni}\}/\lambda(A_{ni}), \tag{2}$$

which, in turn, is expected to approximate $f(x)$ well when the partitions are fine. The size of the partitions has to be chosen by the practitioner. In particular, in the cubic histogram each partition A_{nj} is of the type $\Pi_{i=1}^{r} [a_i k_i h, a_i(k_i + 1)h]$, where a_i are positive constants, k_i are integers and h is a positive parameter. Then $\lambda(A_{nj}) = h^r \Pi_{i=1}^{r} a_i$. The parameter h, called the *binwidth*, controls the hypervolume of the partition and thus the smoothness of the histogram. As h decreases, the number of peaks in the histogram tends to increase. A small h produces estimates with smaller bias but greater variance than estimates based on large h. This trade off between degree of smoothing, bias and variance is shared by all nonparametric curve estimates.

It is clear that the choice of smoothing will affect the shape of the histogram. The histogram competes with more sophisticated nonparametric density estimates because it is easy to compute and is available in many econometric packages. However, some undesirable properties of the histogram are not shared by other nonparametric estimates. The choice of the origin may greatly affect the shape of the histogram (Silverman 1986 Ch. 1 provides several examples). The choice of the coordinate directions in multiple dimensions (i.e. the a_i's) also affects the histogram estimates. Contours are difficult to draw in one dimension. The discontinuity of the histogram prevents estimation of derivatives; these are useful by their own sake, and as intermediate tools in various statistical procedures. Finally, the asymptotic rate of convergence of the histogram to the true density, according to different measures, can be better for alternative density estimates (see section 2.1 below).

The *frequency polygon* smooths out the block-line shape of the histogram by connecting the middle points of each class in the histogram. Another possibility consists of using the histogram method with different 'bins' centred about the point to be estimated. Rosenblatt (1956) proposed the *naive* estimate, for r = 1,

$$\hat{f}(x) = (\hat{F}(x + h) - \hat{F}(x - h))/2h \tag{3}$$

where $\hat{F}(x) = n^{-1} \Sigma_j 1(X_j \leqslant x)$ is the empirical distribution function. The density estimate in (3) is proportional to the relative frequency of the data in an interval of length h and centred about x. The length of the interval has the same role as the bindwidth number and is called a *bandwidth*. Since $\hat{F}(x)$ is an unbiased estimate of $F(x)$ with good statistical properties, the population distribution function, (3) is expected to be a good estimate of $f(x) = dF(x)/dx$ as $h \to 0$.

When r > 1, the naive estimator is defined as

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} 1(X_i \in S(x, h))/\lambda(S(x, h)), \tag{4}$$

where $S(x, h)$ is the sphere in $\mathbb{R}^r$ centred in x and with radius h and $\lambda(.)$ denotes Lebesgue's measure. The numerator in (4) is an unbiased estimator of $\Pr(X \in S(x, h))$, and by Lebesgue's theorem

$$\lim_{h \downarrow 0} \Pr(X \in S(x, h))/\lambda(S(x, h)) = f(x). \tag{5}$$

Then, it is expected that the estimate in (4) will approximate the true density $f(x)$ well for large sample sizes and h small. Note that

$$\begin{aligned} \hat{f}(x) &= (nh^r)^{-1} \sum_i 1(\| x - X_i \| h^{-1} \leqslant 1)/\lambda(S(0, 1)) \\ &= (nh^r)^{-1} \sum_i K((x - X_i)h^{-1}), \end{aligned} \tag{6}$$

where $\| . \|$ is the Euclidean norm and $K(t) = 1(\| t \| \leqslant 1)/\lambda(S(0, 1))$ is the rectangular density on the r-dimensional sphere. Rosenblatt (1956) also suggested smoothing the estimate (4) by replacing the rectangular window in (6), which is a uniform density, by a general function $K: \mathbb{R}^r \to \mathbb{R}$ such that $\int_{\mathbb{R}^r} K(u)\, du = 1$, called a kernel. When K(.) is a density, the resulting density estimate is also a density. In such a case, (6) can be seen as a mixture of densities where each component has the same weight n^{-1}. Popular kernels are:

Gaussian: $K(t) = (2\pi)^{-r/2} \exp(-t't/2)$,

Epanechnikov: $K(t) = \lambda(S(0, 1))^{-1}(r + 2)\{1 - t't)/2\}1(t't < 1)$,

where t is a r-dimensional vector and the prime indicates transposition.

The derivatives of the density are estimated by the derivatives of the density estimate assuming a suitable smooth kernel is used (the Gaussian produces derivative estimates of any order, the uniform, none). That is, if $f^{(s)}(x)$ is the s-th derivative of f at scalar x, it is estimated by

$$\hat{f}^{(s)}(x) = n^{-1} \sum_{i=1}^{n} K^{(s)}(h^{-1}(x - X_i))/h^{r+s},$$

where $K^{(s)}$ is the s-th derivative of the kernel.

Instead of using a scalar bandwidth, one can use a matrix of bandwidths when $r > 1$, in order to take into account the correlation between the components in the vector X, i.e.

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K(H^{-1}(x - X_i))/\det(H), \tag{7}$$

where H is a positive definite matrix. Then (6) is a particular case of (7) with $H = \operatorname{diag}(h, \ldots, h)$. Sometimes, it may be convenient to scale the observations by the sample covariance matrix $\hat{\Sigma}$ as suggested by Fukunaga (1972). That is, use $H = h\hat{\Sigma}$ in (7). Robinson (1983) pointed out that with diagonal bandwidths in a time series context '...the estimated distribution would suggest a "whiter" X_i than is the case, unless X_i is truly white noise'. Kernel density estimates are straightforward to program, but they do entail heavy computation, especially when n is large, the density is to be computed at many points x, and estimates

for several choices of bandwidth are to be found. However, $\hat{f}(.)$ has a convolution form and may thus be computed by using the Fast Fourier Transform algorithm. Silverman (1982) developed this idea for $r = 1$. To estimate the density at a point x, this method requires only the order of $n \log n$ computations, instead of an order of n^2 operations using the naive approach, and not all the computations have to be redone when a new bandwidth is used.

The histograms and kernel estimates discussed above are not locally sensitive to such peculiarities in the data as sparsity in the tails of the density. If the smoothing parameter (binwidth or bandwidth) is too small, values in the tails of the data will provoke bumps or modes in the resulting density estimates. If the smoothing parameter is made too large, in order to avoid such effects, the resulting estimate will be oversmoothed and all the modes may be ironed even in the central part of the density. Several methods implement a local adaptive smoothing where the smoothing parameter is allowed to vary with the data.

An alternative to the histogram estimate (1) is the *variable partition histogram* suggested by Anderson (1965) and Van Ryzin (1973) . When $r = 1$, the partitions depend on the order statistics $X_{(1)}, \ldots, X_{(n)}$. An integer $m \in [2, n]$ is chosen to be the number of bins in the histogram. Then set $k = [n/m]$ ([.] indicates nearest integer) and define partitions $A_{1n} = [X_{(1)}, X_{(k)}]$. $A_{2n} = [X_{(k)}, X_{(2k)}], \ldots,$ $A_{mn} = [X_{((m-1)k)}, X_{(n)}]$. Each partition is of different width and contains about k data points. The density is estimated by (1) using the new partitions. The generalization to multidimensions has been studied by Gessaman (1970).

The basic idea of variable histogram estimates can be also applied to kernels. Fix and Hodges (1951) and Loftsgaarden and Quesenberry (1965) noted that (5) can be approximated in another way, by making the radius of the sphere depend on the sparsity of the data around x, i.e.

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} 1(X_i \in S(x, R_k))/\lambda(S(x, R_k)), \tag{8}$$

where R_k is the minimal value for which $\Sigma_{i=1}^{n} 1(X_i \in S(x, R_k)) = k$. That is, the smoothing parameter is the Euclidian distance between x and its k-th nearest neighbour in the data set. This is called a *nearest neighbour* estimate. Like the naive estimate, the estimate in (8) is expected to perform well for large sample sizes when $R_k \to 0$ (i.e. $k \to \infty$). Friedman *et al.* (1975) proposed an efficient algorithm for locating nearest neighbours. The algorithm has been implemented by Delgado (1990b). Note that (8) can be written as

$$\hat{f}(x) = k/(n\lambda(S(0, 1))R_k^r).$$

The estimated density is not a density itself since $\int_{\mathbb{R}^r} \hat{f}(x)\, dx = \infty$. Mack and Rosenblatt (1979) proposed to make (8) smoother by using a general kernel, i.e.

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K(R_k^{-1}(x - X_i))/R_k^r. \tag{9}$$

Then (8) is a particular case of (9) where K is the uniform density on the unit sphere, and is known as *uniform nearest neighbour* estimate. The estimate (9) is the locally adaptive smooth counterpart of (6), where a different h depending on x is used at each point. The nearest neighbour estimate provides fatter but smoother tails than the kernel.

The *variable kernel* estimate proposed by Breiman *et al.* (1977) adapts the amount of smoothing to the local density of the data avoiding the problem of density estimates with infinite mass under the tails, presented by the nearest neighbours. The variable kernel estimate is defined as

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K(hd_{i,k})^{-1}(x - X_i))/(d_{i,k}H)^r, \tag{10}$$

where $d_{i,k}$ is the Euclidean distance from X_i to its k-th nearest neighbour. Note that the same smoothing is used for each x. Breiman *et al.* (1977) proposed a two step procedure for estimating the window width. In a first step, a pilot estimate $\hat{f}(x)$ is computed using kernels or nearest neighbours such that $\hat{f}(X_i) > 0$ for all i. In a second step, *local bandwidth factors* $\lambda_i = \{\hat{f}(X_i)/g\}^{-\alpha}$ are defined, where $g = \exp(n^{-1} \Sigma_i \log \hat{f}(X_i))$ and $\alpha \in [0, 1]$ is a sensitivity parameter. The density is estimated by

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K(h\lambda_i)^{-1}(x - X_i))/(\lambda_i h)^r.$$

This is known as an *adaptive kernel* estimate. Note that a bandwidth has to be chosen for the pilot estimate as well as the sensitivity parameter α. Breiman *et al.* (1977) recommended using a nearest neighbour estimate as the pilot estimate with a very large value of k and setting $\alpha = 1/r$. Abramson (1982) found that the method is quite insensitive to the pilot estimate and that $\alpha = 1/2$ is a good choice.

Other methods not discussed here are *orthogonal series* (e.g. Fourier series, Laguerre series, Legendre series and Hermite series) and *maximum likelihood* (e.g. convolution sieves and penalized maximum likelihood). These methods are discussed by Prakasa Rao (1983), Devroye and Györfi (1985) and Silverman (1986).

Whittle (1958) found that many nonparametric density estimates can be expressed in term of a *delta sequence* $\{\delta_m(x, y),\ m > 0\}$ which satisfies the condition

$$\int \delta_m(x, y)\phi(y)\, dy \to \phi(x) \quad \text{as} \quad m \to \infty,$$

for every infinitely differential function ϕ with compact support. Then, many density estimators can be written in the form

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} \delta_m(x, X_i).$$

For instance, for the kernel estimate $\delta_m(x, y) = m^r K(m(x - y))$.

2.2. *Asymptotic properties and automatic choice of the smoothing number*

Rosenblatt (1956) noted that all the estimates of the density function satisfying relatively mild regularity conditions are biased. Then, he evaluated the asymptotic mean square error (MSE) of kernel estimates.

The naive estimator in (3) is asymptotically unbiased when $h = h_n$ is a function of the sample size n and

$$h \to 0 \quad \text{as} \quad n \to \infty. \tag{11}$$

The empirical distribution function, $n\hat{F}(x)$, is distributed as a binomial with parameters n and $F(x)$. Therefore

$$E(\hat{f}(x)) = (F(x + h) - F(x - h))/2h \to f(x) \text{ under (11).}$$

Rosenblatt (1956) noted that, assuming the first three derivatives of f(.) exist,

$$F(x + h) - F(x - h) = 2hf(x) + f^{(2)}(x)h^3/3 + 0(h^4).$$

Since

$$MSE = E\{(\hat{f}(x) - f(x))^2\} = f(x)/2h + f^{(2)}(x)^2h^4/36 + 0((nh)^{-1} + h^4).$$

then squared mean consistency is provided by (11) and

$$nh \to \infty \quad \text{as} \quad n \to \infty. \tag{12}$$

Rosenblatt (1956) also proposed general kernel estimates and gave an expression for the asymptotic MSE. The asymptotic properties of the kernel estimate were studied in detail by Parzen (1962). Cacoullos (1966) studied the asymptotic properties of kernel estimates for multivariate densities in (5). We discuss results for the general kernel estimate (7) following Robinson (1983).

Bochner's theorem (Bochner 1955) is used for proving many asymptotic results in kernel estimation. The generalization of this theorem to a general kernel with a matrix of bandwidths H is as follows

Bornech's Theorem. — Let q: $\mathbb{R}^r \to \mathbb{R}$ *be a Borel function such that*

(i) q(.) is continuous at x,

(ii) $\int_{\mathbb{R}^r} |q(u)|\, du < \infty$,

(iii) $\int_{\mathbb{R}^r} K(u)\, du = \chi < \infty$,

(iv) $\int_{\mathbb{R}^r} |K(u)|\, du < \infty$,

(v) $\lim_{\|u\| \to \infty} \|u\|\, |K(u)| = 0$,

(vi) $\overline{\lim}_{n \to \infty} \| H \|^r \det(H)^{-1} < \infty$,

(vii) $\lim_{n \to \infty} H = 0$,

Then

$$\hat{q}(x) = \det(H)^{-1} \int K(H^{-1}(x - u))q(u)\, du \to q(x) \int K(u)\, du \quad as \quad n \to \infty.$$

The conditions of this theorem are very mild. Note that condition (vi) holds when H is a diagonal matrix with equal components.

Then asymptotic unbiasedness follows under the conditions in the theorem, taking $q(u) = f(u)$ and $\chi = 1$, since

$$\hat{q}(x) = E(\hat{f}(x)) = \det(H)^{-1} \int K(H^{-1}(x - u))f(u)\, du \to f(x) \quad as \quad n \to \infty.$$

Consistency follows assuming also that

(viii) $\int_{\mathbb{R}^r} K(u)^2\, du = \alpha < \infty$,

(ix) $n \det(H) \to \infty$.

Note that under independence

$$\text{Var}\,(\hat{f}(x)) = (n \det(H))^{-1}\hat{q}(x) + n^{-1}(E(\hat{f}(x)))^2$$

where

$$\hat{q}(x) = (\det(H))^{-1} \int K(H^{-1}(x - u))^2 f(u)\, du \to f(x)\alpha \quad as \quad n \to \infty.$$

When the density function is smooth, the bias rate of convergence can be improved by using higher order kernels proposed by Bartlett (1963). For simplicity consider the case $r = 1$. We say that a kernel K is of class 0 if it belongs to the class of symmetric kernels about zero which integrate to 1. A *class s kernel* is a class 0 kernel for which

$$\text{(a)} \int |t|^s |K(t)|\, dt < \infty, \quad \text{(b)} \int t^l K(t)\, dt = 0,\ l = 1, \ldots, s - 1. \qquad (13)$$

In view of the symmetry of the kernels, (b) automatically holds for all even values of $l < s$. Most class 0 kernels are class 2 kernels. The order of a class 0 kernel is the largest integer such that K(.) obeys (13). Then if

(x) f(.) is s times boundedly differentiable,
(xi) K(.) is of order s,
(xii) $\sup_u (1 + |u|^2)\, |K(u)| > \infty$,

then $E(\hat{f}(x) - f(x)) = 0(h^s)$. Robinson (1988) provided formulae for constructing higher order kernels of any degree. This bias reduction technique is crucial in many semiparametric procedures. The density estimate can be negative for

kernels of order $s > 2$. This undesirable feature is not so important in semiparametric estimation where nonparametric estimates are only used as an intermediate tool. Applying a Taylor expansion, we obtain an expression for the mean integrated squared error (MISE)

$$\text{MISE} = (s!)^{-2}\beta_s^2 h^{2s} \int f^{(s)}(x)^2 \, dx + n^{-1}h^{-1}\alpha + o((nh)^{-1} + h^{2s}), \qquad (14)$$

where $\beta_s = \int t^s K(t)\, dt$. The MISE is a measure of the discrepancy between the estimated and the true density. Then the MISE depends on the bandwidth, the kernel, and the smoothness of the density through the term $\int f^{(s)}(x)^2 \, dx$, known as *difficulty factor*. From (14), the bandwidth number which minimizes, the asymptotic expansion of the MISE is

$$h_{\text{opt}} \cong \left(\frac{\alpha(s!)^2}{2s\beta_s^2 \int f^{(s)}(x)^2 \, dx} \right)^{1/(2s+1)} n^{-1/(2s+1)}. \qquad (15)$$

Substituting this expression in (14) we obtain the optimal rate of convergence in L_2

$$\inf_h \text{MISE} = 0(n^{-2s/(2s+1)}).$$

Devroye and Györfy (1985) obtained the optimal rate of convergence of the mean integrated absolute error (MIAE)

$$\text{MIAE} = E\left\{ \int_{-\infty}^{\infty} |\hat{f}(x) - f(x)| \, dx \right\}.$$

This is a more robust measure for comparing densities than the MISE. The optimal rate of convergence of the MIAE is $0(n^{-2/(2s+1)})$.

Using calculus of variations, Epanechnikov (1969) proved that the kernel which minimizes α, subject to K(.) a bounded and even density and $\beta_2 = 1$, is the Epanechnikov kernel. The choice of kernel is not crucial. Values of α for different kernels are pretty close. For instance, the ratio of the α's corresponding to the Epanechnikov and Gaussian kernel is 1.051. The choice of the bandwidth number is more important.

Optimal bandwidths can be 'estimated' from the data using (15). Assuming f(.) belongs to a particular parametric family of densities (e.g. the Gaussian). That is, the parameters of the density are estimated (e.g. by maximum likelihood or by a robust procedure), under the assumed parametric model, and then β_s and α are computed by numerical integration. This parametric method has been proposed by Deheuvels (1977) and Deheuvels and Hominal (1980). Suppose we use the Epanechnikov kernel and we assume that $f(x) = f(x, \theta)$ where $\theta = (\mu, \sigma, \gamma)$ is a vector of parameters, μ is the location parameter, σ is the scale parameter and γ is a vector of shape parameters. Then

$$h_{\text{opt}} \approx D_2(\theta)^{1/5} n^{-1/5}, \; D_2(\theta) = 15\left(\int f^{(2)}(x)^2 \, dx \right)^{-1}.$$

It is not necessary to mention location and scale parameters since

$$D_2(\mu, \sigma, \gamma)^{1/5} = \sigma D_2(0, 1, \gamma)^{1/5}.$$

Let $\hat{\sigma}$ be a robust estimate of σ (e.g. the least absolute deviation (LAD) estimate or any other robust estimate of the scale). Then the optimal bandwidths are estimated by

$$\hat{h}_{opt} \approx \hat{\sigma} D_2(0, 1, \hat{\gamma})^{1/5}.$$

where $\hat{\gamma}$ is an estimate of γ computed by any method. Some densities do not have shape parameters (e.g. unimodal densities) and, therefore, it is only necessary to compute the scale parameter from the data. For instance with the normal density and using the Epanechnikov kernel

$$\hat{h}_{opt} \approx 2.345 \hat{\sigma} n^{-1/5}.$$

In practice one may report graphs of the density estimates, computing the optimal bandwidth based on different parametric densities.

Woodroofe (1970) proposed to avoid parameterizing the density in (15) and (22) by estimating it from a preliminary bandwidth, i.e.

$$\hat{h}_{opt} = D_2(\hat{f}, h_0)^{1/5} n^{-1/5}, \qquad D_2(\hat{f}, h_0) = 15 \left(\int \hat{f}^{(2)}(x)^2 \, dx \right)^{-1}.$$

$\hat{f}(x)$ is computed from a given bandwidth h_0. This method is also not completely automatic since h_0 has to be determined by the practitioner. An alternative, suggested by Scott, Tapia and Thompson (1977) is to use an iterative procedure, estimating $\hat{h}_{opt}$ as the solution to

$$\hat{h}_{opt} = D_2(\hat{f}, \hat{h}_{opt})^{1/5}.$$

These methods exclude a large number of densities for which $\int f^{(2)}(x)^2 \, dx$ is not defined or is infinite.

An alternative is to treat h as a parameter which is estimated by optimising some criterion function. For instance Duin (1976) and Habbema *et al.* (1974) proposed choosing h to maximize the cross-validated likelihood

$$L(h) = \prod_{i=1}^{n} \hat{f}_{(-i)}(X_i) \text{ where } \hat{f}_{(-i)}(x) = (n-1)^{-1} \sum_{\substack{j=1 \\ j \neq i}} K(h^{-1}(x - X_j)/h^r.$$

The cross-validation (i.e. the exclusion of the own observation) is due to the fact that $\Pi_{i=1}^{n} \hat{f}(X_i)$ is always maximized for $h = 0$. Chow *et al.* (1983) proved the consistency of this cross-validated estimate assuming f(.) has compact support. However Schuster and Gregory (1981) have found that consistency may not be possible when the density does not have a compact support. However, we can always make a suitable transformation of X such that the corresponding density is defined on a compact. Once, the density of the transformation has been estimated we can obtain the density estimate of X. Simulation studies (see e.g.

Scott and Factor 1981) have shown that the maximum likelihood cross-validation is very sensitive to outliers and to the form of K.

Hall (1983a,b), Rudemo (1982) and Bowman (1982) suggested using the cross-validated estimated mean squared error as criterion function. Then the criterion function is

$$M(h) = \int \hat{f}(x)^2 \, dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{(-i)}(X_i).$$

Stone (1984) found that the h minimizing this function is the best, in the sense of minimizing the MISE.

A central limit theorem for $\hat{f}(x)$ is useful in practice for constructing confidence intervals. For simplicity, consider the case $r = 1$ and a kernel of order 2. Under conditions mentioned above

$$\mathrm{Cov}((nh)^{1/2}\hat{f}(x), (nh)^{1/2}\hat{f}(y)) \to 0 \text{ as } n \to \infty,$$

and for $\{x_1, \ldots, x_s\}$ fixed

$$(nh)^{1/2}(\hat{f}(x_1) - E(\hat{f}(x_1)), \ldots, \hat{f}(x_s) - E(\hat{f}(x_s))) \xrightarrow{d} N(0, V\alpha),$$

where $V = \mathrm{diag}(f(x_1), \ldots, f(x_s))$.

Since

$$(nh)^{1/2}E(\hat{f}(x) - f(x)) = 0((nh^5)^{1/2}),$$

the asymptotic distribution of $(nh)^{1/2}(\hat{f}(x) - f(x))$ is not centred at zero. Assuming that

$$nh^5 \to \gamma < \infty \text{ as } n \to \infty,$$

then, for $\{x_1, \ldots, x_s\}$ fixed points

$$(nh)^{1/2}(\hat{f}(x_1) - f(x_1), \ldots, \hat{f}(x_s) - f(x_s)) \xrightarrow{d} N(\gamma^{1/2}\beta_2 B/2, V\alpha),$$

where $B = (f^{(2)}(x_1), \ldots, f^{(2)}(x_s))'$. The bias disappears when $\gamma = 0$. Note that when $\gamma = 0$, the bandwidth is not optimal. This result can be used for constructing confidence intervals, since V is consistently estimated by

$$\hat{V} = \mathrm{diag}(\hat{f}(x_1), \ldots, \hat{f}(x_s)).$$

Interestingly, the density is estimated more imprecisely on regions where the mass is concentrated.

The asymptotic variance is the same for general kernels (7), but the normalizing factor is $(n \det(H))^{1/2}$. Asymptotic results are not affected under weak dependence. In particular, the asymptotic variance of $\hat{f}(x)$ is unaffected when X_i are strong mixing (Robinson 1983).

Optimal rates of convergence of density estimates have been established by Stone (1980).

Uniform consistency for kernel estimates has been established by Bertrand-Retaly (1978). Devroye and Györfi (1985) have proved that (11) and (12) are

necessary and sufficient conditions for $\int |\hat{f}(x) - f(x)| dx$ to converge to zero with probability one, for any f(.), only assuming that the kernel is a non-negative function which integrates to one.

The consistency of the histogram was established by Rèvèsz (1972). Freedman and Diaconis (1981a) have studied the uniform convergence properties. The histogram has L_2 optimal rate of convergence $n^{-2/3}$ (Scott 1979) and the optimal L_1 rate is $n^{-1/3}$ (Devroye and Györfi 1985). The histogram can be adjusted to enjoy a faster rate of convergence $n^{-4/5}$ in L_2 by smoothing out the block-line shape of the histogram. It has been done using the average shifted histogram (Scott 1985a) and the *frequency polygon* (Scott 1985b).

The consistency of nearest neighbour estimates has been established by Loftsgaarden and Quesenberry (1965) and Moore and Yackel (1977) assuming that

$$k/n \to 0 \text{ and } k \to \infty \text{ as } n \to \infty.$$

L_2 properties have been studied by Rosenblatt (1979) and Mack and Rosenblatt (1979). Devroye and Györfi (1985) noted that it is impossible to study the L_1 properties of nearest neighbour estimates because of their infinite integral. Devroye and Györfi (1985) gave L_1 results for the variable histogram. Similar results for the L_2 case can be found in Prakasa Rao (1983), Lecoutre (1986) and Kogure (1987). The optimal MISE is of the same order as that of the classical histogram. Devroye and Györfi (1985) gave conditions for L_1 consistency of the variable kernel estimate for all f and Devroye and Penrod (1986) proved strong uniform consistency.

Nonparametric kernel density estimates for weakly dependent processes have been studied, for example, by Roussas (1969), Rosenblatt (1971), Robinson (1983), Yakovitz (1985), Roussas (1988), Tran (1989), Györfi *et al.* (1989), Hart and Vieu (1990), and Silveira (1990) and Robinson (1987d) and Hall and Hart (1989) considered strongly dependent processes. The asymptotic properties of the histogram under weak serial dependence have been studied by Györfi (1987) and Tran (1991).

2.3. *Applications*

Density estimates are recommended as a first step in *exploratory data analysis*. Density estimates of residuals from parametric regression may be useful as a preliminary specification tool. Robinson (1987b) considered density estimates of innovations in time series models, based on estimated residuals.

Fix and Hodges (1951) were the first to propose using nonparametric density estimates in application to *discriminatory data analysis*. In discrimination, density estimates $\hat{f}_A(.)$ and $\hat{f}_B(.)$ are computed from samples from two different populations A and B. Then a particular observation $\mathfrak{F}$, is assigned to the population A if $\hat{f}_A(\mathfrak{F}) > \hat{f}_B(\mathfrak{F})$. This method has been applied in medicine in order to perform diagnosis on a disease. The variables are random vectors

containing dummies, which indicate the performance of diagnostic tests, and certain individual characteristics of the patient. One envisages applications of these methods to investigate the success of employment training programs or other social experiments.

In *cluster analysis* the problem is to divide the data set into clusters or classes. Roughly speaking, the problem now is to find from how many and which populations the observations are coming. Several cluster and discrimination methods, using nonparametric density estimation, were reviewed by Prakasa Rao (1983) and Silverman (1986).

Devroye and Györfi (1985) proposed several algorithms for *simulation* from the estimated density and any standard uniform random number generator. They discussed inversion, rejection and order statistics methods for densities estimated from kernels or the histogram. They showed that quite large sample sizes are required for generating observations undistinguishable from observations generated from the true density. These simulation methods are useful in constructing bootstrap estimates. The classical bootstrap performs random sampling with replacement from the empirical distribution of the data. A *smooth bootstrap* consists of generating observations from the estimated density.

Another application is in *testing unimodality*. Silverman (1981, 1983) used the idea that as the amount of smoothing increases, the number of modes or bumps in the estimated density tends to decrease. Thus Silverman proposed finding a critical bandwidth, h_{crit}, such that the nonparametric density estimate is unimodal for any $h \geqslant h_{crit}$. If the true density is unimodal one expects a small h_{crit} and for multimodal densities a large h_{crit}. The h_{crit} is found by a grid search. This idea forms the basis of creating statistics based on h_{crit}. Silverman (1983) proved that as $n \to \infty$, $h_{crit} \to 0$ when the density is unimodal but h_{crit} is bounded away from zero otherwise. The h_{crit} is assessed against a standard family of unimodal densities. Silverman also suggested avoiding use of a parametric family by using the smoothed bootstrap. A large number of samples, with the original sample size, are generated from the estimated density $\hat{f}_{crit}$ computed with h_{crit}. For each replicated sample, new densities are estimated using h_{crit}. Then the proportion of replications producing multimodal densities is the p-value. Craig (1991) has compared these methods with the *dip* tests of Hartigan and Hartigan (1985) in the context of a test for unimodality of fixed costs of labour adjustment. The dip statistic is the maximum difference between the empirical distribution function, and the unimodal distribution function that minimizes the maximum difference. The dip test is asymptotically larger for the uniform distribution than for any distribution in a larger class of distributions.

Tests of independence can also be performed using nonparametric density estimates. Traditionally tests for serial dependence are based on the sample serial correlation or serial rank-based correlation procedures (e.g. Spearman's test). These tests may not be suitable in the analysis of stock market prices, where the dependence may be nonlinear (e.g. ARCH or bilinear processes). Such subtle alternatives to the independence hypothesis may be detected using nonparametric

density estimates. The null hypothesis for testing independence between continuous random variables X and Y can be expressed as H_0: $f(x, y) = f(x)f(y)$ for all x and y. Robinson (1991) proposed using the Kullback and Leibler distance, i.e.

$$I = \int\int f(x, y)\{\log f(x, y) - \log(f(x)f(y))\}\, dx\, dy.$$

This is approximated using kernel estimates of the joint and marginal densities. For example, in order to test serial independence in time series, i.e. independence of X_i and X_{i+1}, a possible statistic is,

$$\hat{I}_n = n^{-1} \sum_i c_i\{\log(\hat{f}(X_i, X_{i+1}) - 2\log(\hat{f}(X_i))\},$$

for a sequence of weights c_i. Robinson (1991) showed consistency of the test based on $\hat{I}_n$ against strong mixing alternatives, and for suitable c_i, asymptotic normality under the null. He applied the resulting test in testing independence of exchange rates for different currencies using daily, weekly and monthly data. Chad and Tran (1992) proposed using the L_1 distance rather than the Kullback–Leibler distance and used histogram instead of kernel estimates. Their test resembles tests based on contingency tables. They showed that the test is consistent when the series is absolutely regular. However, they did not obtain the asymptotic distribution of the statistic under the null and therefore the implementation of their test relies on the estimation of critical values using permutations of the original series. Neither Robinson (1991) or Chan and Tran (1991) considered data dependent bandwidths or binwidths in their theory.

Cumulative distribution functions can be estimated using kernel estimates. In particular the cumulative distribution $F(x)$ is estimated by $\hat{F}(x) = n^{-1} \Sigma_i \mathcal{K}(x - X_i)$, where $\mathcal{K}(.)$ is the cumulative distribution of the kernel (see Prakasa Rao 1983). One may prefer to use the empirical cumulative distribution $\hat{F}(x) = n^{-1} \Sigma_i 1(X_i \leqslant x)$, which is unbiased, avoids the choice of a smoothing parameter, and is computationally very convenient. However, the empirical cumulative distribution may be unsuitable for estimating probabilities under the tails of the density where data are scarce. In such cases a smooth estimator may be preferred. Other functionals, like $f^{(s)}(x)$, $\int f^{(s)}(x)\, dx$ or $\int f(x)^2\, dx$, can be estimated by substituting for the true density the estimated one inside the integrals.

The hazard rate $H(x) = f(x)/(1 - F(x))$ can be estimated by $\hat{H}(x) = \hat{f}(x)/(1 - \hat{F}(x))$. Relevant surveys are Singpurwalla and Wong (1983) and Hassani *et al.* (1986).

The conditional density $f(y \mid X = x) = f(y, x)/f(x)$ is estimated by $\hat{f}_{Y|X}(y) = \hat{f}(y, x)/\hat{f}(x)$ (Watson 1964, Nadaraya 1964 and Rosenblatt 1969).

Other important applications of density estimates are to the computation of regression functions, which will be covered in the next section, and semiparametric estimates, which will be covered in section 4.

3. Nonparametric regression

Econometric models describe the relationship between economic variables. This relationship is often represented by means of conditional moments. In particular, given a $r \times 1$ vector of explanatory variables X, a $q \times 1$ vector of response variables Y and a known function g: $\mathbb{R}^q \to \mathbb{R}$, one is interested in estimating

$$m(x) = E\{g(Y) \mid X = x\} \tag{17}$$

The function m(.) is called a *regression curve*.

Regression curves can be estimated by parametric methods. That is, $m(x)$ is assumed to follow a particular parametric form (e.g. $m(x) = x'\beta$, where β is a vector of unknown parameters) and then the parameters are estimated using some loss function. Economic models usually provide information on some features of the regression curve. For instance, it is known that, for normal goods, the regression curve of expenditure with respect to income has positive first derivatives and negative second derivatives. However, many competing functional forms can be in agreement with economic theory principles.

Nonparametric regression estimates do not impose a rigid functional form on the regression function. Given a data set $\{(Y_i, X_i), i = 1, \ldots, n\}$, the nonparametric estimate of $m(x)$ is a weighted average of $g(Y_i)$, where the heavier weights are given to the observations with X_i closest to x; i.e. $m(x)$ is estimated by

$$\hat{m}(x) = \sum_i g(Y_i) W_{ni}(x), \tag{18}$$

where $\{W_{ni}(x), i = 1, \ldots, n\}$ is a sequence of weights which sum up to one. The idea is that $g(Y_i)$'s with X_i's close to x possess more information on $m(x)$ than observations far away from x. Therefore, $W_{ni}(x)$ will be small when X_i is far away from x. When $W_{ni}(x) = n^{-1}$ all i, (18) is a consistent estimate of $E(g(Y))$ but an inconsistent estimate of $m(x)$.

Nonparametric regression may be used in explanatory data analysis. Hildenbrand and Hildenbrand (1980), Härdle and Jerison (1988), Härdle (1990) and Bierens and Pott-Buter (1991) have obtained nonparametric estimates of Engle curves for different goods. Nonparametric predictors from time series have been also used to investigate the forecastability of rates of return of gold by Prescott and Stengos (1988) and Härdle and Vieu (1989) and of exchange rates by Diebold and Nason (1989). Nonparametric estimates constructed from parametric residuals may be useful in model specification, e.g. to check structure in the residuals or the presence of heteroskedasticity. Estimates of derivatives of the regression function are computed in the same way as density derivatives. In particular, the s-th derivative of $m(x)$, $m^{(s)}(x)$, is estimated by

$$\hat{m}^{(s)}(x) = \sum_i g(Y_i) W_{ni}^{(s)}(x) \tag{19}$$

where $W_{ni}^{(s)}(x) = \partial\, W_{ni}(x)/\partial x$, when the weight function is s-times differentiable. From (19), one can obtain estimates of elasticities and other functionals. Nonparametric regression has also been used in many semiparametric problems. Recent surveys on nonparametric regression are Collomb (1981 and 1985) and books on the topic are Györfi *et al.* (1989) and Härdle (1990).

We next present some specific nonparametric regression techniques, then we discuss automatic choices of the smoothing parameter, and finally we discuss some applications.

3.1. *Some techniques*

Nonparametric density estimates can be represented as a sum of weights, i.e.

$$\hat{f}(x) = \sum_i w(x, X_i).$$

Noting that

$$m(x) = \int g(y) f(g(y) \mid X = x)\, dy = \int g(y) f(g(y), x)\, dy / f(x)$$

$$= P_{g(Y)}(x)/f(x),$$

the problem is to find an estimate of $P_{g(Y)}(x)$. When $g(Y) = 1$, it follows that $P_1(x) = f(x)$ and is estimated by $\hat{f}(x)$. Then, it seems sensible to estimate $P_{g(Y)}(x)$ by

$$\hat{P}_{g(Y)}(x) = \sum_i g(Y_i) w(x, X_i). \tag{20}$$

Then it is possible to construct weights

$$W_{ni}(x) = w(x, X_i) / \sum_i w(x, X_i). \tag{21}$$

So, the histogram produces partition estimates with weights

$$W_{ni}(x) = 1(X_i \in A_{nj}) \left(\sum_{i=1}^{n} 1(X_i \in A_{nj}) \right)^{-1} \text{ when } x \in A_{nj}. \tag{22}$$

Then the partition estimate is just the arithmetic average of the Y_i's with corresponding X_i's in the same bin as x. These type of weights were proposed by McMurtry and Fu (1966), Hill (1969) and Jarvis (1970) among others. Since $\Sigma_{i=1}^{n} 1(X_1 \in A_{nj})$ can be equal to zero for some partitions, it has been suggested that (22) might be replaced by $W_{ni}(x) = n^{-1}$ when $\Sigma_{i=1}^{n} 1(X_i \in A_{nj}) = 0$ for $x \in A_{nj}$ (see Györfi (1990) and Györfi *et al.* (1989)). *Variable partition estimates* are constructed in the same way but using the variable histogram. The *kernel estimates* use weights

$$W_{ni}(x) = K((x - X_i)/h) / \sum_i K((x - X_i)/h), \tag{23}$$

with $W_{ni}(x) = 0$ in the case 0/0. These weights were proposed by Nadaraya (1964) and Watson (1964). More general weights using kernels like those employed in (7) are

$$W_{ni}(x) = K(H^{-1}(x - X_i))/\sum_i K(H^{-1}(x - X_i)). \tag{24}$$

These weights were introduced by Robinson (1983).

The *nearest neighbour estimates* use weights

$$W_{ni}(x) = K((x - X_i)/R_k)/\sum_i K((x - X_i)/R_k). \tag{25}$$

When K(.) is the uniform kernel we have weights,

$$W_{ni}(x) = 1(X_i \text{ is one of the k nearest neighbours of } x)/k. \tag{26}$$

These weights were introduced by Royall (1966), Cover (1968) and Cover and Hart (1967), and weights (25) were introduced by Collomb (1980).

Stone (1977) and Devroye (1978) introduced a general class of nearest neighbour weights. Since the individual coordinates are usually measured in dissimilar units, Stone proposed transforming them to the unit metric before applying the Euclidean metric. The scales used have to satisfy certain conditions. These conditions are met by the sample standard deviation, provided that X admits a nondegenerate distribution and has finite second moments. Then the random (pseudo) metric corresponding to the scales $\{s_{nj}, j = 1, \ldots, r\}$ is defined by

$$\rho_n(u, v) = \left\{\sum_{i=1}^{r} (s_{ni}^{-1}(u_i - v_i)^2\right\}^{1/2},$$

where $u = (u_1, \ldots, u_r)$ and $v = (v_1, \ldots, v_r)$. Let c_{ni} be such that,

$$c_{1n} \geqslant \ldots \geqslant c_{nn} \geqslant 0, c_{ni} = 0 \text{ for } 1 > n \text{ and } \sum_{i=1}^{n} c_{ni} = 1.$$

For $1 \leqslant i \leqslant n$,

$$W_{ni}(x) = \frac{c_{n\nu_i}(x) + \ldots + c_{n\nu_i}(x) + \lambda_i(x) - 1}{\lambda_i(x)} \tag{27}$$

where 1(A) is the indicator function of the event A and

$$\nu_i(x) = (1 + \#(l: l \neq i, \rho_n(X_l, x) < \rho_n(X_i, x)))$$

and

$$\lambda_i(x) = (1 + \#(l: l \neq i, \rho_n(X_l, x) = \rho_n(X_i, x))).$$

This tie breaking rule is computationally expensive. Devroye (1978) suggested breaking the ties by comparing indices; i.e. when X_i and X_j are equally close to x, according to the defined metric, X_i is said to be closer to x if $i < j$. Then, the weights are just $W_{ni}(x) = c_{n\nu_i}$. Stone (1977) required the weights (27) to

satisfy, for asymptotic theory,

$$\text{(i) } c_{n1} \to 0 \text{ and (ii) } \sum_{i \geqslant n\alpha} \to 0 \text{ for all } \alpha > 0 \text{ as } n \to \infty. \tag{28}$$

Devroye (1978) required (28) (i) and that there exists a sequence of numbers $k = k_n$ such that, in asymptotic theory,

$$k \to \infty, k/n \to 0 \text{ and } \sum_{i=k+1}^{n} c_{ni} \to 0 \text{ as } n \to \infty. \tag{29}$$

Devroye (1982) proved that (28) and (29) are equivalent. These properties are satisfied by the following weights:

Uniform: $c_{ni} = k^{-1}$ when $i \leqslant k$ and $c_{ni} = 0$ otherwise, with $k \to \infty$ and $k/n \to 0$ as $n \to \infty$,

Triangular: $c_{ni} = 2(k - i + 1)/(k + k^2)$ when $i \leqslant k$ and $c_{ni} = 0$ otherwise, with $k \to \infty$ and $k/n \to 0$ as $n \to \infty$,

Exponential: $c_{ni} = h(1 + h)^{-1}(1 - (1 - h)^{-n})^{-1}$, with $h \to 0$ and $nh \to \infty$.

In order to check that these weights satisfy (29), Devroye (1978) recommended taking $k \simeq (n/h)^{1/2}$ in (29).

Since $m(x) = P_{g(y)}(x)/f(x)$, one can exploit any available information on $f(x)$, e.g. when $f(x)$ is known (which is highly unlikely in practice), $m(x)$ can be estimated by $\hat{m}(x) = \hat{P}_{g(y)}(x)/f(x)$ (Johnston 1982).

In the fixed design model, where the explanatory variables are non random and possibly equispaced on the interval $[0, 1]$, Priestley and Chao (1972) proposed weights (with $r = 1$)

$$W_{ni}(x) = (X_i - X_{i-1})h^{-1}K((x - X_i)/h).$$

Gasser and Müller (1979) defined weights for the fixed design model

$$W_{ni}(x) = h^{-1} \int_{S_{i-1}}^{S_i} K((x - u)/h)\, du, \tag{30}$$

where $X_{i-1} \leqslant S_{i-1} \leqslant X_i$ is chosen between the ordered X's. Cheng and Lin (1981) considered the case $S_i = X_i$.

Yang (1981) and Stute (1984) considered a type of nearest neighbour estimate with weights

$$W_{ni}(x) = n^{-1}h^{-1}K((\hat{F}(x) - \hat{F}(X_i))/h), \tag{31}$$

where $\hat{F}(.)$ is the empirical cumulative distribution. These weights are of the form (23) where we use the fact that $\Sigma_i n^{-1}h^{-1}K((\hat{F}(x) - \hat{F}(X_i))/h) \to 1$, under the usual conditions on h.

Nonparametric regression in a time series context has been studied by, Watson (1964), Roussas (1969), Bosq (1980), Robinson (1983), Doukhan and Ghindès (1980 and 1983), Collomb (1984), Yakowitz (1985 and 1987) and Bierens (1990) among others.

Note that the weights in (18) can be interpreted as minimizing a square loss function i.e.

$$\hat{m}(x) = \text{argmin}_{\mu} \sum_j (g(Y_j) - \mu)^2 W_{nj}(x).$$

Then $\hat{m}(x)$ is highly sensitive to the effect of just one isolated observation Y_i, particularly if the corresponding X_i is close to x. The idea of robust estimation of a location parameter has been adapted to this context. In particular, the influence of outlying observations is reduced by substituting for the quadratic loss function a convex function $\rho(.)$ with bounded derivative $\phi(.)$. That is, the robust conditional location functional $r(x)$ defined by the equation

$$E\{\phi(g(Y) - r(x)) \mid X = x\} = 0,$$

is estimated by the solution to

$$\sum_j \phi(g(Y_j) - r(x)) W_{ni}(x) = 0. \tag{32}$$

This estimator was proposed by Tsybakov (1983), Robinson (1984) and Härdle (1984) using kernels. Härdle and Tsybakov (1988) extended it to simultaneous robust estimation of conditional location and scale using kernels and Boente and Fraiman (1989, 1990) to location invariant robust estimates, estimating the conditional scale *a priori*, using kernel and nearest neighbour estimates.

Conditional quartiles can be estimated from the estimated cumulative conditional distribution $F(\mathcal{G} \mid X = x)$,

$$\hat{F}(\mathcal{G} \mid X = x) = \sum_j 1(g(Y_j) \leqslant \mathcal{G}) W_{nj}(x). \tag{33}$$

Then the α-th conditional quartile ζ_α is estimated by $\hat{\zeta}_\alpha$ such that

$$\hat{F}(\hat{\zeta}_\alpha \mid X = x) = \alpha. \tag{34}$$

These quartile estimates were defined by Stone (1977) using general consistent weights. He also proposed L-estimates computed by functions of conditional quartile estimates.

Stone (1977) also defined *local linear weights* as follows. Let us define $Z_i' = (1, X_i')$, then $\hat{\beta} = \{\Sigma_i Z_i Z_i' W_{ni}(x)\}^{-1} \Sigma_i Z_i g(Y_i) W_{ni}(x)$ minimizes the function

$$\sum_j (g(Y_j) - \beta' Z_j)^2 W_{nj}(x).$$

The weights $V_{ni}(x) = x' \{\Sigma_i Z_i Z_i' W_{ni}(x)\}^{-1} \Sigma_i Z_i W_{ni}(x)$ are called local linear weights. These weights may be inconsistent and Stone proposed a transformation of $V_{ni}(x)$ which produces consistent weights. Cleveland (1979), Cleveland and Devlin (1988) and Cleveland *et al.* (1988) proposed local linear polynomial weights which are robustified by means of an iterative procedure.

3.2. *Asymptotic properties and automatic choice of the smoothing number*

The general kernel estimate using weights (24) can be expressed as

$$\hat{m}(x) = \hat{P}_{g(y)}(x)/\hat{f}(x),$$

where

$$\hat{P}_{g(y)}(x) = (n \det(H))^{-1} \sum_j g(Y_j)K(H^{-1}(x - X_j)),$$

estimates

$$P_{g(y)}(x) = m(x)f(x).$$

The asymptotic unbiasedness of $\hat{P}_{g(y)}(x)$ follows from Bochner's theorem, taking $q(u) = P_{g(y)}(u)$, $f(.)$ and $m(.)$ are continuous at x, and $E \mid m(X) \mid < \infty$. Then

$$\begin{aligned}\hat{q}(u) = E(\hat{P}_{g(y)}(x)) &= \det(H)^{-1}E(g(Y)K(H^{-1}(x - X))) \\ &= \det(H)^{-1}E(m(X)K(H^{-1}(x - X))) \\ &= \det(H)^{-1} \int P_{g(y)}(U)K(H^{-1}(x - u))\, du \\ &\to P_{g(y)}(x) \text{ as } n \to \infty.\end{aligned}$$

The asymptotic bias of $\hat{P}_{g(y)}(x)$ can be reduced, when $P_{g(y)}(x)$ admits enough derivatives, by using higher order kernels, as in density estimation. In particular, using kernels of order two and assuming also that the first two derivatives of $P_{g(y)}(x)$ exist and are bounded and (xii), when $r = 1$,

$$E(\hat{P}_{g(y)}(x) - P_{g(y)}(x)) \approx 2^{-1}h^2\beta_2\, \partial^2 P_{g(y)}(x)/\partial x^2.$$

Mean square consistency follows assuming also (vii)–(ix) and $s^2(x) = E(g(Y)^2 \mid X = x)$ is continuous at x, since assuming independence

$$\mathrm{Var}(\hat{P}_{g(y)}(x)) = (n \det(H))^{-1}\hat{Q}(x) + n^{-1}E(\hat{P}_{g(y)}(x)),$$

where

$$\hat{Q}(x) = (\det(H))^{-1} \int s(u)^2 f(u)K(H^{-1}(x - u))\, du \to s(x)^2 f(x)\alpha \text{ as } n \to \infty.$$

Therefore, $\hat{m}(x)$ is also consistent, applying Slutsky's theorem. Strictly, the moments of kernel estimates may only exist under restrictive conditions in view of the random denominator $\hat{f}(x)$.

The sequence of weights $\{W_{ni}(x), i \geqslant 1\}$ is said to be *universally L_s consistent* for $s \geqslant 1$ fixed, if

$$E\left\{ \int \mid \hat{m}(x) - m(x) \mid^s f(x)\, dx \right\} \to 0 \text{ as } n \to \infty,$$

for all possible distributions of (Y, X) such that $E \mid g(Y) \mid^s < \infty$.

Stone (1977) gave necessary and sufficient conditions for the universal consistency of weights and he applied his result to prove the universal consistency of nearest neighbour weights. Devroye and Wagner (1980) and Spiegelman and Sacks (1980) proved the L_s universal consistency of kernel weights and Györfi (1990) the universal consistency of partition estimates. An important feature of many of these results is that, despite the motivation given for kernel estimates, X need not have a density, so that discrete valued regressors are permitted.

Schuster (1972) established for $r = 1$, under conditions stated above, and

$$nh^3 \to \infty,\ nh^5 \to 0 \text{ as } n \to \infty,$$

for $\{x_1, \ldots, x_s\}$ distinct fixed points,

$$(nh)^{1/2}(\hat{m}(x_1) - m(x_1), \ldots, \hat{m}(x_s) - m(x_s)) \xrightarrow{d} N(0, \alpha W),$$

where $W = \text{diag}\{\sigma^2(x_1)/f(x_1), \ldots, \sigma^2(x_s)/f(x_s)\}$ and $\sigma^2(x) = \text{Var}(Y \mid X = x)$. The conditional variance is consistently estimated by

$$\hat{\sigma}^2(x) = \sum_i Y_i^2 W_{ni}(x) - \left\{\sum_i Y_i W_{ni}(x)\right\}^2.$$

Hence,

$$\hat{W} = \text{diag}\{\hat{\sigma}^2(x_1)/\hat{f}(x_1), \ldots, \hat{\sigma}^2(x_s)/\hat{f}(x_s)\}$$

is a consistent estimate of W that can be used for constructing consistent confidence intervals. Robinson (1983) proved that for $r > 1$ and using weights (24), the asymptotic variance is the same, but the normalizing factor is $(n \det(H))^{1/2}$. He also proved that the asymptotic variance is unaffected under weak dependence where the regressors are lagged values of the dependent variable and assuming that the series is strong mixing.

The plug-in method is more difficult to implement than in density estimation. The MSE of $\hat{m}(x)$ may not exist because it is a ratio between random variables. The following linearization is often applied

$$\hat{m}(x) - m(x) \approx (\hat{P}_{g(y)}(x) - m(x)\hat{f}(x)))/f(x).$$

Then, the MSE of $\hat{m}(x)$ is approximated by

$$E(\hat{P}_{g(y)}(x) - m(x)\hat{f}(x)))^2/f(x)^2 \approx (nh)^{-1}\sigma^2(x)\alpha f(x)^{-1} + h^4\{\beta_2(m^{(2)}(x) + 2m^{(1)}(x)(\partial \log f(x)/\partial x))/2\}^2,$$

where $m^{(i)}(x) = \partial^i m(x)/\partial x^i$. The 'optimal bandwidth' which minimizes the above expression is proportional to $n^{-1/5}$ and depends on the unknowns $\sigma^2(x)$, $f(x)$, $f^{(1)}(x)$, $m^{(1)}(x)$, $m^{(2)}(x)$. All these unknowns must be estimated when implementing the plug-in procedure.

Alternatively, other measures of accuracy can be employed. The mean integrated weighted squared error (MIWSE) is defined as

$$E\left\{\int \{m(x) - \hat{m}(x)\}^2 f(x) V(x)\, dx\right\}, \tag{36}$$

where $V(.)$ is a weighting function. Stone (1982) provided the lower and optimal rate of convergence in this sense, under certain smoothness conditions on m(.) and f(.).

The MIWSE can be estimated from the data by the averaged weighted squared error (AWSE)

$$\text{AWSE} = n^{-1} \sum_j \{g(Y_j) - \hat{m}(X_j)\}^2 V(X_j).$$

The bandwidth minimizing this function will be too small since $\hat{m}(X_j) \to g(Y_j)$ as $h \to 0$, for fixed n. Clark (1975) proposed the leave-one-out cross-validation function

$$\text{CV}(h) = \sum_j \{g(Y_j) - \hat{m}_{(j)}(X_j)\}^2 V(X_j), \tag{37}$$

$\hat{m}_{(j)}(x) = \Sigma_{i \neq j} g(Y_i) W_{ni}(x)$ and $W_{ni}(x)$ are kernel weights.

An alternative to the leave-one-out cross-validation function is to use some penalizing function. In this case h is chosen to be $\hat{h} = \text{argmin}_h Q(h)$, where

$$Q(h) = \sum_j \{g(Y_j) - \hat{m}(X_j)\}^2 \Psi(n^{-1}h^{-1}) V(X_j),$$

where $\Psi(.)$ is the penalizing function which corrects for too small h. Examples of penalizing functions are:

$\Psi(n^{-1}h^{-1}) = (1 - n^{-1}h^{-1}K(0))^{-2}$, Craven and Wahba (1979) and Li (1985).

$\Psi(n^{-1}h^{-1}) = \exp(2n^{-1}h^{-1}K(0))$, Akaike (1970).

$\Psi(n^{-1}h^{-1}) = (1 + n^{-1}h^{-1}K(0))/(1 - n^{-1}h^{-1}K(0))$, Akaike (1974).

$\Psi(n^{-1}h^{-1}) = (1 - 2n^{-1}h^{-1}K(0))^{-1}$, Rice (1984).

Härdle and Marron (1985a and 1985b) have proved that, under certain regularity conditions, all these data-driven bandwidths are asymptotically optimal with respect to different accuracy measures. Consistent cross-validated smoothing parameters in the nearest neighbour case have been obtained by Li (1984). Liz (1987) proved the asymptotic optimality of several cross-validation criteria where the smoothing parameter takes discrete values, with application to nearest neighbours estimation. Andrews (1991a) has generalized his result to models with heteroskedastic disturbances.

3.3 *Applications*

Nonparametric regression estimates can be employed to check the goodness of particular parameterizations of conditional expectations and the usefulness of candidate explanatory variables without specifying functional form. Plots of parametric and nonparametric fitting are useful in this respect.

Bierens and Pott-Butter (1990) demonstrated the usefulness of nonparametric regression analysis for functional specification of household Engle curves.

Household demand functions and equivalence scales are estimated from econometric models where the demand function is specified in advance, directly or indirectly, via the specification of the utility function or cost function. The functional form of the model is chosen on the basis of tractability rather than on the basis of prior knowledge of the true functional form. The probability of choosing the correct model is small due to the many functional forms theoretically admissible. Misspecification of functional form leads to inconsistent parameter estimates and the equivalence scales will also be inconsistent estimated. Bierens and Pott-Butter (1990) noted that the life cycle consumption hypothesis leads to demand systems that relate specific demand to net income, prices and household composition, plus a heteroskedastic error term. The demand functions involved are conditional expectation functions. The functional form of the Engle curve is estimated by nonparametric kernel regression using the 1980 Budget Survey for the Netherlands. Two Engle curves are estimated for expenditure on food, clothing and footwear, and for other expenditures. The regressors are net income, number of children in the age group 0–15, and in the age group 16 or over. They plotted the Engle curves of expenditure versus net income for the different household categories. Then, parametric models were chosen in accordance with the nonparametric regression results. The final specifications of the Engle curves are linear, depending on income and the number of children in the two age groups.

Stone (1977) proposed estimates of conditional variances, covariances and correlation functions. In particular the conditional variance of Y given the vector of regressors X, $\sigma^2(x) = \mathrm{Var}(Y \mid X = x)$, is estimated by

$$\hat{\sigma}^2(x) = \sum_i Y_i^2 W_{ni}(x) - \left\{ \sum_i Y_i W_{ni}(x) \right\}^2. \tag{38}$$

Rose (1978) proposed other conditional variance estimates in the linear regression model when it is known that $\mathrm{E}(Y \mid X = x) = x'\beta$, where β is a vector of unknown parameters. A possible estimate of $\sigma^2(x)$ is

$$\hat{\sigma}^2(x) = \sum_j (Y_i - X_i'\hat{\beta})^2 W_{ni}(x), \tag{39}$$

where $\hat{\beta}$ is some preliminary estimate of β; e.g. the least squares estimate.

Elasticity and other economic functional estimates can be computed from the estimates of the derivatives of the regression curve.

Another important application is in prediction of time series. The regression function in this case is an autoregressive process of unknown functional form, i.e. let $Z_i' = (X_{i-1}, X_{i-2}, \ldots, X_{i-p})$ and

$$m(\mathfrak{F}) = \mathrm{E}\{X \mid Z = \mathfrak{F}\}.$$

In efficient asset markets, it is widely agreed that high frequency asset returns are linearly unpredictable, conditionally heteroskedastic and conditionally leptokurtic. Empirical and theoretical results are consistent with the conjecture that nonlinearities may be present in asset returns' conditional means. However,

the out-of-sample forecasting of linear models has not been improved on by any nonlinear model. Diebold and Nason (1990) provided several explanations for this fact: nonlinearities may be present in even-ordered conditional moments, nonlinearities such as outliers may be present, and it is difficult to find the correct parametric model. They estimated the conditional expectation of ten major dollar spot rates using the nonparametric local weighted regression proposed by Cleveland (1979) and Cleveland and Devlin (1988). They found that nonparametric regression does not improve the out-of-sample prediction of a simple random walk. These results are consistent with those of Presscott and Stengos (1988) who were unable to improve forecast of gold prices in the Canadian market using kernel regression.

Nonparametric estimates of conditional variances in a time series context are useful for estimating the stock return volatility in asset markets. Pagan and Schwert (1990) compared the nonparametric conditional stock volatility estimates with parametric estimates like the autoregressive conditionally heteroskedastic (ARCH) model, generalized ARCH (GARCH), exponential GARCH, and Markov's switching-regime.

An important application consists of testing the difference between regression curves. Suppose that we have data $\{(X_i, Y_i, Z_i),\ 1 \leqslant i \leqslant n\}$ from the random variable (X, Y, Z). We want to test the hypothesis H_0: $m_y = m_z$ where $m_y(x) = E(Y \mid X = x)$ and $m_z(x) = E(Z \mid X = x)$. King (1989) proposed the statistic

$$\hat{\tau}_n = n^{-1} \sum_i (\hat{m}_y(X_i) - \hat{m}_z(X_i))^2,$$

where $\hat{m}_y(.)$ and $\hat{m}_z(.)$ are kernel estimates which employ the same bandwidth. He obtained the small sample distribution of such a statistic on the assumption that the error terms in each regression are normally distributed, and the asymptotic distribution of $\hat{\tau}_n$, after suitable normalization, when the bandwidth tends to zero as $n \to \infty$. Härdle and Marron (1990) obtained a similar test where the difference between the two regressions is parameterized. Hall and Hart (1990) proposed a statistic

$$\hat{S}_n = \left[\sum_{j=0}^{n-1} \left(\sum_{i=j+1}^{j+m} D_i\right)^2\right] \left[n \sum_{j=0}^{n-1} (D_{i-1} - D_i)^2/2\right]^{-1},$$

where m is the integer part of nh for fixed h, and $D_i = Y_i - Z_i$ for $1 \leqslant i \leqslant n$, and $D_i = D_{i-n}$ for $n + 1 \leqslant i \leqslant n + m$. They proved, keeping h fixed, that $\hat{S}_n$ provides an asymptotically powerful test, and the asymptotic distribution of $\hat{S}_n$, under the null, is a Wiener process. They proposed to determine the critical values by bootstrap, which are more accurate than the critical values obtained by the asymptotic approximation. Hall and Hart (1990) generalized this test to the case of testing several nonparametric functions, and regression curves which depend on different regressors.

Many other applications of nonparametric regression are found in semiparametric problems which are discussed in the next section.

4. Estimation of semiparametric models

Many if not most econometric models are semiparametric. A parametric structure explaining some basic economic phenomena (e.g. utility or cost functions) is usually known and one is interested in the estimation of these parameters and in making inferences on the assumed structure from the data. However, many features of the data generating process are of unknown form. The full functional form of the distribution cannot usually be justified from economic theory and nor is it likely to be of specific economic interest. In the recent econometric literature, estimation of a number of semiparametric models requires nonparametric estimation of certain functionals in the first step. In these models, it is explicitly recognized that certain features of the underlying distribution of the data are unknown while others follow a known parametric model. The goal is to obtain estimates for the parametric part that are asymptotically equivalent to those obtained when the nonparametric part of the model is perfectly known.

In nonparametric estimation, 'nature' only provides a data set $\{Z_i, i \geqslant 1\}$, and different features of the underlying data generating process are estimated from these data. In semiparametric estimation, 'nature' also provides some parametric relationship between variables. For instance,

$$E(Y \mid X = x) = x'\beta, \tag{41}$$

$$E(Y \mid X = x, Z = z) = x'\beta + \mu(z) \text{ with } \mu \text{ unknown}, \tag{42}$$

$$\text{Median}(Y \mid X = x, W = w) = x'\beta. \tag{43}$$

In (41) and (43), the parameters of interest β, can be consistently estimated by parametric methods, e.g. least squares or robust estimation. The problem in this case is to choose the most appropriate objective function. Semiparametric estimates improve the efficiency of simple consistent parametric estimates by incorporating estimates of some unknown nonparametric function in the objective function. Consistent estimates are sometimes impossible to obtain by parametric methods (e.g. in (42)) but they can be obtained by semiparametric estimation.

Thus, in a semiparametric problem one combines a known function of the parameter of interest (e.g. $x'\beta$), with a nonparametric shape function G, such as the density of the disturbances, regression functions, conditional quantiles etc. A semiparametric estimate is adaptive if it is asymptotically as efficient as the infeasible estimate which employs a correct finite parameterization of G. (In what follows when referring to 'efficiency', we shall always mean 'asymptotic efficiency'). The property of adaptation refers to the existence of such an estimate. Unfortunately, adaptation is not always possible.

Stein (1956) proved that a necessary condition for adaptation is that for every finite parameterization G_η of G, where η is a finite dimensional vector, the limiting covariance matrix of the infeasible estimate of β and the nuisance parameters vector η is block diagonal. It implies that knowledge of η cannot

improve estimation of β. Bickel (1982) gave a condition for a less general class of G's, having the heuristic interpretation that the efficient estimate based on a given G is still root-n-consistent when G is misspecified. Manski (1984) gave a necessary condition for adaptation of a subset of parameters in β and Schick (1986) gave a condition that is necessary in models more general than Bickel's.

Begun *et al.* (1983) showed how to obtain lower bounds for the asymptotic covariance matrix of estimates of β when adaptation is not possible. These bounds, called semiparametric efficiency bounds, have been calculated for certain econometric models, see Chamberlain (1986, 1990) and Cosslett (1987) and Newey (1990).

In order to conserve on space and to achieve a more unified presentation we focus on methods which employ the smoothed nonparametric density or regression estimates introduced in sections 2 and 3. Thus we omit such important semiparametric methods as least absolute deviations (LAD) estimates of censored regression (where nonparametric estimation occurs only in computing standard errors) and maximum score estimates of discrete choice models (which are asymptotically non-normal and are not root-n-consistent). Some of these methods were included in the survey of Robinson (1988b). Not a great deal of work has been done on the choice of data dependent bandwidth in semiparametric problems, though of course the rules described in sections 2 and 3 can be applied.

In the rest of this section we present different semiparametric estimates. Surveys of the semiparametric literature are Robinson (1988b), Newey (1990b) and Härdle (1990). Most semiparametric work has employed the kernel or nearest neighbour nonparametric estimates discussed above, including the bulk of the work referred to below. Another semiparametric method which is proving popular in a variety of semiparametric problems is series estimation, see e.g. Andrews (1991b) and Newey (1991).

4.1. *Full adaptive estimation*

Bickel (1982) considered the linear regression model (41). Under regularity conditions, the Cramer–Rao efficiency bound is achieved by the one step score estimate

$$\bar{\beta} = \tilde{\beta} - \left\{ \sum_i X_i X_i' f^{(1)}(U_i)^2 f(U_i)^{-2} \right\}^{-1} \sum_i X_i f^{(1)}(U_i) f(U_i)^{-1},$$

where $\tilde{\beta}$ is a preliminary root-n-consistent estimate, and $U_i = Y_i - \beta' X_i$. However, the density of the disturbances U_i is nonparametric and the disturbances themselves are as unobserved. Stone (1975), for the case $X = 1$ and β scalar, and Bickel (1982), for the general regression case, suggested an estimate of approximately the form

$$\hat{\beta} = \tilde{\beta} - \left\{ \sum_i X_i X_i' \hat{f}^{(1)}(\tilde{U}_i)^2 \hat{f}(\tilde{U}_i)^{-2} \right\}^{-1} \sum_i X_i \hat{f}^{(1)}(\tilde{U}_i) \hat{f}(\tilde{U}_i)^{-1}, \tag{44}$$

where $\tilde{U}_i = Y_i - \tilde{\beta}' X_i$ and $\hat{f}(\tilde{U}_i)$ and $\hat{f}^{(1)}(\tilde{U}_i)$ are kernel estimates of f(.) and $f^{(1)}(.)$. Bickel proved that $\hat{\beta}$ is as efficient as the infeasible $\bar{\beta}$, assuming that the U_i are iid, symmetric and independent of X_i. Symmetry is not necessary for adaptive estimation of the slope coefficient. Bickel required splitting of the sample into two parts. The residuals $\tilde{U}_i$ are computed from one part of the sample and the density and its derivatives are estimated from the other. Schick (1986) employed a less drastic form of sample-splitting. Manski (1984) extended Bickel's results to nonlinear regression, avoiding independence between disturbances and regressors. A Monte Carlo study of bandwidth choice is reported in Hsieh and Manski (1987). Kreiss (1987) extended these results to adaptive estimation of the coefficients of an ARMA model. He did not require the sample-splitting of Bickel. Steigerwald (1990) considered an extension to regression models with ARMA errors.

Engle and González-Rivera (1989) have applied this method to the estimation of ARCH models with conditional density of unknown form. They assumed a linear functional form for the conditional expectation and the conditional variance of stock returns. Then, using ordinary least squares, they estimated the parameters of the conditional mean and variance in order to estimate the standardized residuals. The density function of the standardized residuals is estimated by a kernel estimate. Then, the log-likelihood function based on the estimated density is maximized with respect to the parameters of the conditional mean and variance. The information matrix is not block diagonal, and the resulting semiparametric estimate is not proved to be adaptive. However, they reported encouraging Monte-Carlo results.

There are other semiparametric estimates, based on nonparametric estimates of the score function, that achieve certain efficiency bounds. Newey and Powell (1987a and b) used nonparametric score estimates to achieve semiparametric efficiency bounds within a certain class of distributions for censored regression models under symmetry and conditional quantile restrictions. Lee (1990) has used also estimates of the score function in sample selection models. His semiparametric estimate achieves the Chamberlain (1986) bound in the binary selection model.

4.2. *Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form*

Consider model (41) with $\mathrm{Var}(Y \mid X = x) = \sigma^2(x)$ of unknown form. The infeasible weighted least squares estimate

$$\bar{\beta} = \left\{ \sum_i X_i X_i' / \sigma^2(X_i) \right\}^{-1} \sum_i X_i Y_i / \sigma^2(X_i),$$

is Gauss–Markov efficient under suitable regularity conditions. Rose (1978) suggested estimating $\sigma^2(X_i)$ by (38) or (39) and then estimating the coefficients

$$\hat{\beta} = \left\{ \sum_i X_i X_i' / \hat{\sigma}^2(X_i) \right\}^{-1} \sum_i X_i Y_i / \hat{\sigma}^2(X_i). \tag{45}$$

This estimate has been proved to be as efficient as $\bar{\beta}$ by Carroll (1982) using (38) with kernel weights for the single regression model, assuming that $\sigma^2(.)$ is a smooth function and the regressor admits a density with compact support and is bounded away from zero. Robinson (1987a) relaxed Carroll's assumptions to moment conditions in the general multiple regression model using (38) with nearest neighbour weights. He did not require continuity of $\sigma^2(.)$, and indeed allowed X to have a discrete or mixed distribution, not only a continuous one. Lee (1990c) presented Monte Carlo results using a data driven bandwidth. Delgado (1989b) has extended Robinson's results to the multiple equations nonlinear regression model and Delgado (1989a) has proved adaptation in the nonlinear regression model using (39).

A natural application of this method is to count regression models. When Y is a count variable taking values $0, 1, 2, \ldots$, the conditional variance is typically a function of the conditional mean. The usual estimation approach in econometrics is maximum likelihood (ML). Under regularity conditions, the ML estimate is asymptotically efficient if the conditional distribution of Y given X is correctly specified (e.g. a Poisson, a geometric or a negative binomial). Furthermore, when the conditional distribution is incorrectly specified, the pseudo ML (PML) estimate may be consistent but asymptotically inefficient (e.g. the Poisson ML). The usefulness of the Poisson model is limited by the fact that the variance is equal to the conditional mean, which is rarely obviously true using microeconomic data. Other likelihood functions, like the geometric or negative binomial, may yield inconsistent pseudo ML (PML) estimates under misspecification. In this sort of model, it is typically assumed that $E(Y \mid X = x) = g(x, \beta_0)$, where β_0 are unknown parameters and $g(x, \beta)$ is always positive for any value of β; e.g. $g(x, \beta) = \exp(x'\beta)$. The semiparametric weighted nonlinear least squares estimate

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i (Y_i - g(x, \beta))^2 / \hat{\sigma}^2(X_i),$$

is, asymptotically, equally efficient to the Poisson ML estimate when the likelihood is correctly specified, but it is more efficient than the Poisson PML estimate when the conditional distribution is not Poisson. The semiparametric estimate is, in general, more inefficient than other ML estimates when the likelihood is correctly specified. But it is consistent more generally than PML estimates under misspecification.

Delgado and Kniesner (1990) considered the problem of modelling worker absenteeism on London buses. In their application, Y is the number of absence spells of up to 7 days in 1985. X is a vector consisting of: the numbers of absences in 1984 and 1983; and variables including pay grades, nonwage characteristics of employment, worker's personal and economic characteristics,

the employer's technological and economic traits, and the legal and economic environment. There were $n = 5101$ observations, a promising number for semiparametric estimation, because sufficiently good nonparametric conditional variance estimates are likely to be obtainable. It was assumed that $g(\alpha, \beta) = \exp(\alpha'\beta)$, which is common in regression models of counts. They computed semiparametric estimates as well as ordinary least squares and ML (based on various models). It was found that the semiparametric and negative binomial ML results were similar.

Müller and Stadtmüller (1987) obtained efficient estimates in the fixed design model with unknown heteroskedasticity, using weights (38). Harvey and Robinson (1988) considered a nonstationary time series regression model with trending regressors and stationary autoregressive disturbances which are multiplied by an unknown time-varying factor. They obtained efficient estimates of the regression parameters using a Cochrane–Orcutt algorithm where the residuals are standardized by a nonparametric estimate of the heteroskedasticity based on partition weights. Robinson (1986) also used weights (36) to obtain estimates of time varying parameters and heteroskedastic variances which are a function of time.

Robinson (1987b) has proposed applying semiparametric methods to the estimation of models with ARCH effects, where the conditional variance is of unknown form. The conditional variances and their derivatives are estimated by nonparametric regression. Then, estimates of the parameters of the conditional mean are obtained substituting the nonparametric estimates in the normal equations of the Gaussian log-likelihood. Applications of this method may be found in Whistler (1989) and Lee (1990b).

The estimate defined in (45) can be arbitrarily influenced by outlying observations. Delgado (1990a) proposed using the estimates defined in (38) and (47) with nearest neighbour weights to correct for heteroskedasticity in the linear regression model using robust estimates. This estimate is implicitly defined as the solution to

$$\sum_i \Psi((Y_i - X_i'\hat{\beta})/\hat{\sigma}(X_i), X_i)X_i/\hat{\sigma}(X_i) = 0, \tag{46}$$

where $\Psi()$ is a bounded function and $\hat{\sigma}(X_i) = (\hat{\sigma}^2(X_i))^{1/2}$. Delgado proved, under regularity conditions, that this estimate is asymptotically as efficient as the infeasible estimator which employs the true $\sigma(X_i)$ in (46).

4.3. *Optimal semiparametric instrumental variable estimators*

Consider a nonlinear simultaneous equations model expressed by the conditional moment restriction

$$E(U(\beta_0, Y, X) \mid X) = 0, \tag{47}$$

where $U(.)$ is a $q \times 1$ vector of known functions, β is an unknown $p \times 1$ vector of parameters, X is a $r \times 1$ random variable and Y is a $s \times 1$ random variable.

It is also assumed that

$$\mathrm{Var}(U(\beta_0, Y, X) \mid X = x) = \Omega,$$

where Ω is positive definite and independent of x. Let H_i be a matrix of instruments, such that

$$H_i = R(X_i, \beta_0)\Omega^{-1},$$

where

$$R(X_i, \beta) = E(\partial U(\beta, Y, X)/\partial\beta \mid X = X_i).$$

When U is nonlinear in Y, in general at least some elements of $R(X_i, \beta)$ are of unknown functional form, because (47) produces insufficient information. Amemiya (1977) defined the optimal nonlinear three stage least squares (NL3SLS) estimate as

$$\bar{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i U_i(\beta)' H_i' \left(\sum_i H_i \Omega H_i' \right)^{-1} \sum_i H_i U_i(\beta).$$

Arbitrary instruments uncorrelated with U_i will generally produce root-n-consistent but asymptotically inefficient estimates. Let $\tilde{\beta}$ be some preliminary root-n-consistent estimate of β. When the (α, τ)-th component of $R(\beta, X_i)$, $r_{\alpha\tau}(\beta, X_i)$, is of unknown functional form, it is estimated by

$$\hat{r}_{\alpha\tau}(X_i) = \sum_j c_{\alpha\tau}(Y_j, X_j, \tilde{\beta}) W_{nj}(X_i),$$

where $c_{\alpha\tau}(Y_j, X_j, \beta)$ is the (α, τ)-th component of the matrix $\partial U(\beta, Y_i, X_i)/\partial\beta$. The variance matrix Ω is estimated by its sample analogue

$$\hat{\Omega} = n^{-1} \sum_j U_j(\tilde{\beta}) U_j(\tilde{\beta})'.$$

Then, the optimal instruments H_i are estimated by $\hat{H}_i = \hat{R}_i \hat{\Omega}^{-1}$, where $\hat{R}_i$ has (α, τ)-th component $\hat{r}_{\alpha\tau}(X_i)$; and β would be estimated by

$$\bar{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_i U_i(\beta)' \hat{H}_i' \left(\sum_i \hat{H}_i \hat{\Omega} \hat{H}_i' \right)^{-1} \sum_i \hat{H}_i U_i(\beta).$$

This feasible NL3SLS estimate has been proposed by Newey (1990a). Newey (1990a) has proved, under conditions similar to Robinson (1987), that the corresponding one-step estimate is asymptotically efficient.

If certain independence assumptions are added to (47), Robinson (1990a) has shown that estimates as efficient as Newey's can be obtained by estimating $R(\beta, X_i)$ by resampling techniques. In particular, if

$$c_{\alpha\tau}(Y_i, X_i, \beta) = t_{\alpha\tau}(\xi(\beta), h(\beta_0, Y_i, X_i), X_i),$$

where $t_{\alpha\tau}(.)$, $\xi(.)$ and $h(.)$ are known functions , $V_i = h(\beta_0, Y_i, X_i)$ is independent of X_i. Then,

$$\hat{r}_{\alpha\tau}(X_i) = T^{-1} \sum_{j \in \Im} t_{\alpha\tau}(\xi(\tilde{\beta}), h(\tilde{\beta}, Y_j, X_j,), X_i),$$

where $\mathfrak{I}$ is a subset of size T of the integers $\{1, 2, \ldots, n\}$. Robinson (1990a) proved that the resulting NL3SLS estimate of β is as efficient as the infeasible estimate, under regularity conditions which allow for lagged independent variables or serially correlated disturbances, but requires independence between V and X, often tantamount to independence between U and X, strengthening the conditional moment restriction (47). An advantage is that the estimate avoids the choice of a smoothing parameter. The number of elements in $\mathfrak{I}$ has to increase with n but at an arbitrarily slow rate, where this rate may affect a Berry–Esseen-type bound (that is the rate of convergence to the limiting distribution).

Application of these methods is interesting in transformed regression models, where PML generally yields inconsistent estimates. Consider for instance the transformed model

$$U(\beta, Y_i, X_i) = \operatorname{arcsinh}(\lambda Y_i)/\lambda - \alpha - \theta' X_i,$$

where Y is scalar and $\beta = (\lambda, \alpha, \theta')'$ is the vector of parameters. Then

$$R(X_i, \beta) = (r(X_i), -1, -X_i),$$

where

$$\begin{aligned} r(X_i) &= E\left(\partial U(\beta_0, Y, X)/\partial\lambda \mid X = X_i\right) \\ &= E\left[\tanh(\lambda_0(\alpha_0 + \beta_0' X + U)^2/\lambda_0) - (\alpha_0 + \beta' X)/\lambda_0 \mid X = X_i\right]. \end{aligned}$$

Root-n-consistent estimates of θ are obtained using arbitrary instruments, uncorrelated with U. The resulting estimates $\tilde{\theta} = (\tilde{\alpha}, \tilde{\lambda}, \tilde{\beta}')'$ are inefficient. However, efficient estimates are obtained applying Newey's method, estimating $r(X_i)$ by

$$\hat{r}(X_i) = \sum_j \{\tanh(\tilde{\lambda}(\tilde{\alpha} + \tilde{\beta}' X_j + \tilde{U}_j)^2/\tilde{\lambda}) - (\tilde{\alpha} + \tilde{\beta}' X_j)/\tilde{\lambda}\} W_{nj}(X_i).$$

Assuming X is independent of U, we can apply Robinson's method. The estimate of $r(X_i)$ is given by

$$\hat{r}(X_i) = T^{-1} \sum_{j \in \mathfrak{I}} \{\tanh(\tilde{\lambda}(\tilde{\alpha} + \tilde{\beta}' X_j + \tilde{U}_j)^2/\tilde{\lambda}) - (\tilde{\alpha} + \tilde{\beta}' X_j)/\tilde{\lambda}\}.$$

Note that $V_i = U_i$ in this case.

These methods have been extended and applied to the estimation of price elasticities of demand for car attributes and fuel efficiency by Lee (1990a).

4.4. *Semiparametric partially linear models*

Semiparametric estimates of model (42) have been considered by Spiegelman (1976), Green *et al.* (1985), Engle *et al.* (1986), Rice (1986), Heckman (1986), Robinson (1988a), Speckman (1988), Carroll and Härdle (1989) and others.

Note that

$$E(Y \mid Z = \mathfrak{F}) = E(X \mid Z = \mathfrak{F})'\beta + \gamma(\mathfrak{F}).$$

Then, with serially uncorrelated and homoskedastic errors, an efficient estimate of the slope coefficients is obtained by regressing $Y_i - E(Y \mid Z = Z_i)$ on $X_i - E(X \mid Z = Z_i)$, assuming X and Z are not functionally related. A feasible version can be constructed by estimating the conditional expectations by nonparametric methods. In particular, Robinson (1988a) proposed estimating β by

$$\hat{\beta} = \left\{ \sum_i X_i^* X_i^{*\prime} \hat{I}_i \right\}^{-1} \sum_i X_i^* Y_i^* \hat{I}_i,$$

where $X_i^* = X_i - \hat{m}_X(Z_i)$ and $Y_i^* = Y_i - \hat{m}_Y(Z_i)$, $\hat{m}_X(Z_i)$ and $\hat{m}_Y(Z_i)$ are higher order kernel estimates of $E(X \mid Z = Z_i)$ and $E(Y \mid Z = Z_i)$, respectively, and $\hat{I}_i = 1(\hat{f}(Z_i) > b)$ where $\hat{f}(Z_i)$ is the corresponding density estimate of Z evaluated at Z_i, and b is a small trimming number. Under regularity conditions $\hat{\beta}$ is root-n-consistent and asymptotically normal. Chamberlain (1990) has shown that it achieves a semiparametric efficiency bound. Higher order kernels are used in order to make the bias of $\hat{\beta}$ (see Speckman 1988) converge at the appropriate rate. Lee (1990c) has provided Monte Carlo results using data driven bandwidths.

Lee (1989) has applied this approach to the estimation of the 'surprise' consumption function. The model can be expressed as follows

$$Y_i = \beta'(X_i - E_i(X_i)) - \gamma' E_i(X_i) + U_i,$$

where Y_i is consumption and X_i is a vector observable, $E_i(X_i)$ is an unobservable vector of agents expectations of X_i, and U_i is a scalar unobservable. The vector $X_i - E_i(X_i)$ consists of 'surprises' or 'news'. It is assumed that agents have rational expectations, that is

$$E_i(X_i) = E(X_i \mid \mathfrak{I}_i),$$

where $\mathfrak{I}_i$ is the information set at time i. Further, it is assumed that $\mathfrak{I}_i$ can be summarized by a vector of observables Z_i, so that

$$E(X_i \mid \mathfrak{I}_i) = E(X_i \mid Z_i).$$

It is supposed that $h(\mathcal{F}) = E(X \mid Z = \mathcal{F})$ is nonparametric; we know the information set governing agents' expectations but we do not know by which mechanism these are formed. The consumption function can be described as semiparametric. We can rewrite the model as

$$\begin{aligned} Y_i &= \beta' X_i + (\gamma - \beta)' h(Z_i) + U_i \\ &= \beta' X_i + \varphi(Z_i) + U_i, \end{aligned}$$

where $\varphi(Z_i) = (\gamma - \beta)' h(Z_i)$ is of unknown functional form. This is a partially linear regression model. Robinson (1989) considered an alternative but related class of statistics which provides tests of certain hypothesis (such as $\beta = 0$) under fairly general serially dependent, stationary, observations on X_i, Z_i and Y_i. Lee (1989) used USA quarterly data and considered the model

$$\Delta C_i = \gamma_1 + \gamma_2 E_i(r_i) + \gamma_3 E_i(I_i) + \beta_1(r_i - E_i(r_i)) + \beta_2(I_i - E_i(I_i)) + U_i,$$

where C_i is consumption, $\Delta C_i = C_i - C_{i-1}$, I_i is income, and r_i is interest rate. Lee suggested that the information set $\mathfrak{I}_i$ might initially comprise a large number of variables, such as three lags of C, I, nominal interest rates, averaged hours worked per capita, government expenditures, inflation and stock prices. Because $n = 133$ only, nonparametric estimates of $E_i(r_i)$ and $E_i(I_i)$ when Z_i is a 21-dimensional vector will be hopelessly imprecise, and though the effect on estimates of β_1 and β_2 is likely to be less serious, it is clearly desirable to seek a Z_i which contains much of the information in the 21 variables, but is of much smaller dimension. This is an extremely difficult and delicate task, specially as the possibly nonlinear character of h(.) makes a linear components procedure to be possibly inadequate. Lee employed a principal components procedure based on certain nonlinear functions, as well as linear ones, and found it possible to choose a two-dimensional Z_i Among his conclusions, the semiparametric tests tended to find semiparametric coefficients insignificant, and the estimates of the surprise coefficients tended to differ between parametric and semiparametric methods. The semiparametric estimates were found to be sensitive to the choice of bandwidth number, but not excessively so.

Stock (1989) uses this semiparametric estimate for nonparametric policy analysis. He considers model (42) where observations are drawn from different cells. The variables Z include policy variables that can be modified by the policy maker, and X are dummy variables indicating the cell specific effects. The dependent variable measure the success of the policy. Then, the objective of this research is to estimate the benefit of a particular policy by the average

$$B_n = n^{-1} \sum_j [E(Y \mid X = X_j, Z = Z_j) - E(Y \mid X = X_j, Z = Z_j^*)]$$

$$= n^{-1} \sum_j (\gamma(Z_j) - \gamma(Z_j^*)),$$

where Z_i and Z_i^* are the values of Z before and after the policy intervention. Stock (1989) proposed to estimate B_n by

$$\hat{B}_n = n^{-1} \sum_j (W_{nj}(Z_i) - W_{nj}(Z_i^*))(Y_j - \hat{\beta}' X_j),$$

and obtained the asymptotic distribution of $\hat{B}_n$ under regularity conditions and after suitable normalization. Stock (1991) applied this procedure to the estimation of the mean hazardous waste clean up benefits. In his application, equation (42) is interpreted as an hedonic price equation, where Y is the price of a house, Z are waste related and waste not related housing characteristics. The housing waste related characteristics are proxy variables indicating the risk of the waste site. These variables are a function of the distance from the house to the waste site, the area of the waste site, and whether or not the waste site is hazardous. The not waste related characteristics are the size of the lot, living area in the house, a measure of the neighbour status, the age of the house, and the distance from the house to the centre of the town weighted by the town

population. The data consist of 324 single family homes in eleven western and northwestern Boston suburbs. The dependent variable is the sale price of the house between April 1978 and March 1981 deflated to 1980 prices according to the annual National Association of Realtors. Stock (1991) compared ordinary least squares results, assuming $\gamma(.)$ is linear, and the semiparametric methods. He found that the range of semiparametric benefit estimates was comparable to the range of estimates using least squares. However, both semiparametric and parametric methods resulted in imprecise estimates of the benefit of clean up the hazardous waste site.

The approach of Engle *et al.* (1986) and Heckman (1986) is based on spline estimates. Powell (1989) and Newey *et al.* (1990), considered (42) where $Z_i = W_i'\delta$, W_i are observable variables and δ is an unknown vector parameters with application to censored regression models. In this model δ can be estimated root-n-consistently and then, a similar approach to that discussed above is employed. Ahn and Powell (1990) considered Z_i to be an unobservable regression function which can be estimated nonparametrically.

Semiparametric estimates such us those described in sections 4.1 to 4.4 are included in a general class considered by Andrews (1990). He also explicitly considered hypothesis testing rules in semiparametric models, estimates which avoid sample splitting, and allowance for serial dependence and mild heterogeneity.

4.5. *Semiparametric estimation based on averaged nonparametric estimates and their derivatives*

Powell *et al.* (1989) proposed an estimate of

$$\delta = \mathrm{E}\{\mathrm{f}(\mathrm{X})\mathrm{m}^{(1)}(\mathrm{X})\},$$

where $\mathrm{m}^{(1)}(\mathrm{x})$ are the first derivatives of the unknown regression function $\mathrm{m}(\mathrm{x}) = \mathrm{E}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x})$, and $\mathrm{f}(.)$ is the density of X. It is interesting in a number of econometric applications to limited dependent variable models where we have an index model of the form

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x}) = \mathrm{m}(\mathrm{x}'\beta_0),$$

where β_0 is a $\mathrm{r} \times 1$ vector of unknown parameters. In this case $\delta = \mathrm{c}\beta_0$, where c is an unknown constant. Powell *et al.* noted that, under mild regularity conditions, integration by parts produces

$$\delta = -2\mathrm{E}\{\mathrm{Yf}^{(1)}(\mathrm{X})\}.$$

Thus, $\mathrm{f}^{(1)}(.)$ can be estimated by kernels and δ by

$$\hat{\delta}_\mathrm{n} = -2\mathrm{n}^{-1} \sum_\mathrm{i} \mathrm{Y}_\mathrm{i} \hat{\mathrm{f}}_{(\mathrm{i})}^{(1)}(X_\mathrm{i})$$

$$= \binom{\mathrm{n}}{2}^{-1} \sum_{\mathrm{i}=1}^{\mathrm{n}-1} \sum_{\mathrm{j}=\mathrm{i}+1}^{\mathrm{n}} \mathrm{p}_{\mathrm{ij}},$$

where $p_{ij} = -h^{-r-1}K^{(1)}((X_i - X_j)/h)(Y_i - Y_j)$, $K^{(1)}$ is the vector of first derivatives of the kernel function, and $\hat{f}^{(1)}(.)$ is the derivative of $\hat{f}_{(i)}(.)$ defined in (16). Applying asymptotic theory for U-statistics, they proved that

$$n^{1/2}\,\mathrm{Var}(r(X, Y))^{-1/2}(\hat{\delta}_n - E(\hat{\delta}_n))/2 \xrightarrow{d} N(0, I_r)$$

where

$$r(x, y) = f(x)m^{(1)}(x) - \{y - m(x)\}f^{(1)}(x).$$

A kernel estimate of $r(X_i, Y_i)$ is

$$\hat{r}(X_i, Y_i) = (n-1)^{-1} \sum_{\substack{j \neq 1 \\ j \neq i}}^{n} p_{ij},$$

Since $E(r(X, Y)) = \delta$, a consistent estimate of $\mathrm{Var}(r(X, Y))$ is

$$n^{-1} \sum_i \hat{r}(X_i, Y_i)\hat{r}(X_i, Y_i)' - \hat{\delta}_n\hat{\delta}_n'.$$

Powell *et al.* (1989) noted that $E(\hat{\delta}_n) - \delta = o(n^{-1/2})$ by using high order kernels, and assuming enough derivatives of f(.). They also proposed an instrumental variables estimate of $\delta^* = \delta/E(f(X))$. They noted that

$$\delta^* = \{E(f^{(1)}(X)X')\}^{-1}E(f^{(1)}(X)Y).$$

Then δ^* is estimated by

$$\hat{D}_n = \hat{\delta}_{xn}^{-1}\,\hat{\delta}_n,$$

where $\hat{\delta}_{xn} = \Sigma_i f^{(1)}(X_i)X_i'$.

Härdle and Stoker (1989) have obtained semiparametric estimates of $\delta = E\{\partial g(X)/\partial X\}$, namely $\hat{\delta}_n = -n^{-1}\,\Sigma_i(Y_i\hat{f}^{(1)}(X_i)/\hat{f}(X_i))\hat{I}_i$, where $\hat{I}_i$ was defined in the last section. Newey and Stoker (1990) considered efficiency properties of average derivative estimates. Härdle *et al.* (1989) have studied the MSE error properties of average derivatives estimates. They found that the bandwidth minimizing the mean square error is proportional to $n^{-2/7}$. Stoker (1989) used average derivatives estimates in testing additive derivative constraints. Robinson (1989) used them in testing a variety of hypotheses in parametric and nonparametric time series models. His methods were applied and extended by Lee (1989) in analysis of the 'surprise' consumption function reviewed in section (4.4). Ahn and Manski (1990) used related methods in the analysis of binary choice models with nonparametric estimation of expectations. Samarov (1990) has proposed different tests based on averaged second derivatives.

4.6. *Asymptotically efficient estimation in the presence of autocorrelation of unknown form*

Consider a linear regression model

$$Y_i = X_i'\beta + U_i,\ i = 1, \ldots, n,$$

where the X_i are time series and the U_i are stationary and nonparametrically autocorrelated. The model can be written in vector form as,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U},$$

where $E(\mathbf{UU}') = \Gamma$ is an unknown Toeplitz matrix. Then the infeasible generalized least squares estimate

$$\bar{\beta} = (\mathbf{X}'\Gamma^{-1}\mathbf{X})^{-1}\mathbf{X}'\Gamma^{-1}\mathbf{Y},$$

is Gauss–Markov efficient. The problem is to obtain feasible estimates that are as efficient as $\bar{\beta}$. This problem can be solved by premultiplying the model by the Fourier matrix which diagonalizes Γ. Then the model becomes a regression model with approximately independent heteroskedastic disturbances. The disturbance variances are proportional to spectral density ordinates of U. If the spectral density were known or correctly parameterized, efficient estimates could be computed in a standard way. Hannan (1963) proposed estimating the spectral density nonparametrically, showing that the corresponding frequency-domain generalized least squares estimate is as efficient as the infeasible one under quite general conditions, and allowing for trending regressors. This idea has been used in other econometric models, such as distributed lags (Hannan 1965), linear simultaneous equations (Hannan and Terrell 1973), continuous time systems (Robinson 1976) and regression models with conditional heteroskedasticity of unknown form (Hidalgo 1989). Robinson (1990b) justified efficiency of the estimates with general data driven smoothing parameter for the spectral density, justifying the consistency of a particular choice of smoother.

Hidalgo (1990) has considered the problem of nonlinear autoregressive disturbances, where $U_i = \rho(U_{i-1}) + \varepsilon_i$, $i = 1, \ldots, n$, $\rho(.)$ is an unknown function and ε_i are white noise. He proposed a Cochrane–Orcutt type estimate where $\rho(.)$ is estimated nonparametrically.

4.7. *Linear regression parameter estimation constructed by nonparametric estimation*

Faraldo Roca and González Manteiga (1985), Cristóbal Cristóbal *et al.* (1989) and Stute and González Manteiga (1990) considered the estimation of the linear regression model $Y = X'\beta^0 + \varepsilon$ where ε is independent of X and $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$, by the general class of estimators

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \int (\hat{m}_Y(x) - x'\beta)^2 \, d\Omega_n(x),$$

where $\Omega_n(x)$ is a weighting function and $\hat{m}_Y(x)$ is a nonparametric estimate of $E(Y \mid X = x)$. They propose to use the weighting function,

$$\Omega_n(x) = \int_{-\infty}^{x} \hat{f}(t)\, dt,$$

where $\hat{f}(x)$ is the nonparametric estimate of the density of X evaluated at x.

Faraldo Roca and González Manteiga (1985) and Cristóbal Cristóbal *et al.* (1989) proved that $\hat{\beta}$ is to first order as efficient as ordinary least squares. They showed good performance of the biased $\hat{\beta}$ with respect to ordinary least squares when mean squared error (MSE) is used for comparison. It obviously requires a 'judicious choice' of the smoothing parameter. Faraldo Roca and González Manteiga (1985) calculated the optimal bandwidth using kernels which minimize the MSE of $\hat{\beta}$ in the one regressor case. For this choice of bandwidth the MSE of $\hat{\beta}$ is smaller than the variance of the ordinary least squares.

A related non-iterative estimation method of a linear regression model with censored data has been proposed by González Manteiga and Cadarso Suárez (1990).

4.8. *Testing parametric versus nonparametric hypothesis*

The problem of testing a parametric specification is of considerable importance in econometrics, for example in testing the adequacy of a linear regression model. Traditionally the approach used in econometrics has been of a parametric character, in that the parametric null hypothesis is nested within a parametric alternative; Lagrange multiplier tests have been particularly popular. However, tests against nonparametric alternatives can also be conducted, problems involving comparison between parametric and nonparametric fits.

The central limit theorem discussed in section 3.2 can be used for testing the hypothesis of a linear regression versus a weakly specified nonparametric alternative, as proposed by Robinson (1983). With X scalar, H_0: $m(x) = \alpha + \beta x$, where α and β are unknown parameters. Then under the null $\Phi m = 0$, where Φ is a $(s-2) \times s$ matrix and $m = (m(x_1), \ldots, m(x_s))'$ and $\{x_1, \ldots, x_s\}$ are distinct fixed points. Then, under the null

$$\hat{\tau}_n = nh\hat{m}'\Phi'(\Phi\alpha\hat{W}\Phi')^{-1}\Phi\hat{m} \xrightarrow{p} \chi^2_{s-2},$$

where $\hat{m} = (\hat{m}(x_1), \ldots, \hat{m}(x_s))'$ is the kernel estimate of m, and $\hat{W}$ and α were defined in section 3.1 and 3.2. The statistic $\hat{\tau}_n$ can be used for testing the linearity hypothesis.

Azzalini *et al.* (1989) have proposed a likelihood ratio test for testing the functional form of the conditional mean in count data models. They considered observations of a count variable Y_i taking values $0, 1, 2, \ldots$. The null hypothesis is H_0: $E\{Y \mid X = x\} = x'\beta$ versus weakly specified alternatives of the form H_1: $E\{Y \mid X = x\} = m(x)$, where m(.) is an unknown function. They also assumed that the conditional density of Y given X is such that $f(Y = y \mid X = x) = \mathscr{L}(y, m(x))$, that is, conditional density functions which are completely determined by their first conditional moment, e.g. the binomial or the Poisson. The statistic used is

$$\tau_n = \sum_i \{\log \mathscr{L}(Y_i, X_i'\hat{\beta}) - \log \mathscr{L}(Y_i, \hat{m}(X_i))\},$$

where $\hat{\beta}$ is computed by maximum likelihood and $\hat{m}(X_i)$ uses kernel weights. They applied this test to the case that $\mathscr{L}(.)$ is binomial with parameter $p(\mathfrak{x}) = 1/\{1 + \exp(-\mathfrak{x}'\beta)\}$ under the null, and parameter of unknown functional form under the alternative. They also applied the method when $\mathscr{L}(.)$ is Poisson with mean $\mathfrak{x}'\beta$ under the null and with mean of unknown functional form under the alternative. They also discussed the generalization of the method to the case where the conditional distribution is also a function of additional nuisance parameters η, i.e. $f(Y = \mathcal{y} \mid X = \mathfrak{x}) = \mathscr{L}(\mathcal{y}, m(\mathfrak{x}), \eta)$ and applied the test to a AR(1) model. The implementation of the test is based on bootstrapping in the absence of knowledge of the asymptotic null distribution of the statistic.

Delgado and Stengos (1990a) considered tests for the competing hypothesis

$$H_0\text{: } E(Y \mid X = \mathfrak{x}, Z = \mathfrak{F}) = \mathfrak{x}'\beta_0 \text{ versus } H_A\text{: } E(Y \mid X = \mathfrak{x}, Z = \mathfrak{F}) = m(\mathfrak{F}),$$

where m(.) is unknown and X and Z does not completely overlap. That is, the hypotheses are non-nested. They proposed a Davidson and MacKinnon (1982) type test where the two hypotheses are artificially nested by means of the comprehensive regression model

$$Y_i = X_i'\theta_0 + \delta\hat{m}(Z_i) + \text{error},$$

where $\hat{m}(.)$ is a nearest neighbour regression estimate based on the Z regressors, and θ_0 is a parameter vector. Then, the least squares t-ratios for δ are asymptotically distributed as a standard normal under the null. The least squares estimate of δ, in the above regression, converges in probability to 1 under the alternative. These t-ratios are used for testing H_0 versus H_A. Simulations reported in Delgado and Stengos (1990a) are encouraging. They also consider the case where the model in the null is nonlinear in parameter using J, C, and P type tests.

A J-test procedure based on estimated residuals has been considered by Wooldridge (1990), using sieve estimates. B. Lee (1991) has also proposed a residual specification test based on the residuals from kernel regression. Cox *et al.* (1988) considered generalized spline models for regression. They were concerned with testing that the regression function is of a particular parametric form against the alternative that the function is partially linear (in the sense of section 4.4). Eubank and Spiegelman (1990) proposed an alternative spline based methodology for testing the goodness of fit of a linear model. Yatchew (1990) and Yatchew and Bos (1991) proposed tests for the difference between two partially linear models. Tests using the average derivative method have been studied in Stoker (1989), Robinson (1989), and Samarov (1991).

5. Software

There is often a trade off between computational effort and efficiency. Nonparametric estimates are relatively easy to compute. For instance, in GAUSS or MATLAB, a Gaussian kernel estimate for r = 1, with a bandwidth h, and a data vector stored in the $n \times 1$ array x, is computed at point u by means of the

sentence

$$\text{density} = \text{sum}(\exp(-(u-x).^*(u-x)/(2^*h^*h)))/(\text{sqrt}(2^*\text{pi})^*h^*n),$$

('sum' must be changed by 'sumc' in GAUSS). However, this approach is very inefficient. If we want to obtain estimates at each data point, we can exploit the symmetry of the kernel for reducing the number of computations and the storage size. If we want just to plot the density or regression estimates, we can make use of the Fast Fourier Transform, as suggested by Silverman (1982) and Härdle (1987).

Programs for nonparametric regression are available in abundant supply. The kernel method has been implemented in International Mathematical and Statistical Libraries, Inc. (1984) (IMSL) as subroutine NDKER and IMSL (1987) as subroutine DESKN. In both routines the user provides the kernel function and the bandwidth. The language S (Becker and Chambers 1984), also provides density estimates. The package CURVDAT provides FORTRAN routines for density, regression, density derivatives and regression derivatives.

The package TIMESLAB (Newton 1988) provides kernel density estimates using different kernels. The package XploRe (Broich *et al.* 1990) performs different nonparametric estimation procedures with excellent graphical capabilities (see No and Sickles 1990 and Lee 1992 for reviews of this software). The package N-Kernel (see Delgado and Stengos 1990 and Lee 1992 for reviews of this software) implements a particular method based on local kernel weights which is very useful in investigating departures from linearity in regression. Delgado (1990b) provided a number of FORTRAN routines using kernels and nearest neighbours and discussed their application in solving semiparametric problems using standard econometric software.

Acknowledgements

This article is based on research funded by the Economic and Social Research Council (ESRC) reference number: R000231441.

References

Abramson, I. S. (1982) On bandwidth variation in kernel estimates — a square root law, *Annals of Statistics*, 10, 1217–1223.

Ahn, H. and Manski, C. F. (1990) Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations, Preprint.

Ahn, H. and Powell, J. L. (1990) Semiparametric estimation of censored selection models with a nonparametric selection mechanism, Preprint.

Akaike, H. (1970) Statistical predictor information, *Annals of the Institute of Statistical Mathematics*, 22, 203–217.

—— (1974) A new look at the statistical model identification, *IEEE Transactions of Automatic Control*, AC-19, 716–723.

Amemiya, T. (1977) The maximum likelihood and the nonlinear three-stage least squares estimator in the nonlinear simultaneous equation model, *Econometrica*, 45, 955–968.

Andrews, D. W. K. (1990) Asymptotics for semiparametric econometric models: I estimation and testing, Preprint.

—— (1991a) Asymptotic optimality of generalized C_L, cross-validation and generalized cross-validation in regression with heteroskedastic errors, *Journal of Econometrics*, 47, 359–377.
—— (1991b) Asymptotic normality of series estimators for various nonparametric and semiparametric models, *Econometrica*, (forthcoming).
Anderson, T. W. (1965) Some nonparametric multivariate procedures based on statistically equivalent blocks, in *Multivariate Analysis I*, (ed. P. R. Krishnaiah).
Azzalini, A., Bowman, A. and Härdle, W. (1989) On the use of nonparametric regression for model checking, *Biometrika*, 76, 1–12.
Bartlett, M. S. (1963) Statistical estimation of density functions, *Sankhya A* 25, 145–154.
Bean, S. J. and Tsokos, C. P. (1980) Developments in nonparametric density estimation, *International Statistical Review*, 48, 267–287.
Becker, R. A. and Chambers, J. M. (1984) *S: An interactive environment for data analysis and graphics*, Belmont, CA: Wadsworth.
Begun, J. M., Hall, W. J., Huang, W. M. and Wellner, J. A. (1983) Information and asymptotic efficiency in parametric–nonparametric models, *Annals of Statistics*, 11, 432–452.
Bertrand-Retali, M. (1978) Convergence uniforme d'un estimateur de la densité par la methodé de noyau', *Rev. Roumaine Math. Pures. Appl.*, 23, 361–385.
Bickel, P. (1982) On adaptive estimation, *Annals of Statistics* 447–471.
Bierens, H. J. (1990) Model free asymptotically best forecasting of stationary economic time series, *Econometric Theory*, 6, 348–383.
Bierens, H. J. and Pott-Buter, H. A. (1991) Specification of household Engle curves by nonparametric regression, *Econometric Reviews*, 9, 123–184.
Bochner, S. (1955) *Harmonic Analysis and the Theory of Probability*, Chicago: University of Chicago.
Boente, G. and Fraiman, R. (1989) Robust nonparametric estimation for dependent observations, *Annals of Statistics* 17, 1242–1256.
—— (1990) Asymptotic distribution of robust estimators for nonparametric models from mixing processes, *Annals of Statistics*, 18, 891–906.
Bosq, D. (1980) Une méthode nonparamétrique de prédiction d'un processus stationnaire. Prédictions d'une mesure aléatoire, *C.R. Acad. Sci. Paris, Sér. A.*, 290, 711–713.
Bowman, A. W. (1985) A comparative study of some kernel-based nonparametric density estimators, *Journal of Statistical Computation and Simulation*, 21, 313–327.
Breiman, L., Meisel, W. and Purcell, E. (1977) Variable kernel estimates of multivariate densities, *Technometrics* 19, 135–144.
Broich, T., Härdle, W. and Krause, A. (1990) XploRe — a computing environment for Exploratory Regression and Analysis, Springer-Verlag (forthcoming).
Cacoullos, T. (1966) Estimation of a multivariate density, *Annals of the Institute of Statistical Mathematics*, 18, 178–189.
Carroll, R. J. (1982) Adapting for heteroskedasticity in linear models, *Annals of Statistics* 10, 1224–1233.
Carroll, R. J. and Härdle, W. (1989) A note on second-order effects in a semiparametric context, *Statistics*, 20, 179–186.
Chamberlain, G. (1986) Asymptotic efficiency in semiparametric models with censoring, *Journal of Econometrics*, 32, 189–218.
—— (1990) Efficiency bounds for semiparametric regression, Preprint.
Chan, N. H. and Tran, L. T. (1992) Nonparametric tests for serial dependence, *Journal of Time Series Analysis* (forthcoming).
Cheng, K. F. and Lin, P. E. (1981) Nonparametric estimation of a regression function, *Zeitschrift fur Wahrsheinlichkeitstheorie und verwandte Gebiete*, 57, 223–233.
Chow, Y. S., Geman, S. and Wu, L. D. (1983) Consistent cross-validated density estimation, *Annals of Statistics*, 11, 25–38.
Clark, R. M. (1975) A calibration curve for radiocarbon dates, *Antiquity*, 49, 251–266.

Cleveland, W. S. (1979) Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, 74, 829–836.
Cleveland, W. S. and Devlin, S. J. (1988) Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, 83, 596–610.
Cleveland, W. S., Devlin, S. J. and Grosse, E. (1988) Regression by local fitting: methods, properties and computational algorithms, *Journal of Econometrics*, 37, 87–114.
Collomb, G. (1980) Estimation de la régression par la méthode des k points les plus proches acec noyau: quelques propriétiés de convergence ponctuelle, *Lecture Notes in Mathematics 831*, 159–175.
—— (1981) Estimation non-paramétrique de la régression: revue bibliographique, *International Statistical Review*, 49, 75–93.
—— (1984) Proprietés de convergence presque complète du prédicteur à noyau, *Zeitschrift fur Wahrsheinlichkeitstheorie und verwandte Gebiete*, 66, 441–460.
—— (1985) Non-parametric regression: an up-to-date bibliography, *Statistics*, 16, 309–324.
Cosslett, S. J. (1987) Efficiency bounds for distribution-free estimators of the binary choice and censored models, *Econometrica*, 55, 559–585.
Cover, T. M. (1968) Estimation by the nearest neighbor rule, *IEEE Transactions on Information Theory*, IT-14, 50–55.
Cover, T. M. and Hart, P. E. (1967) Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, IT-13, 21–27.
Craig, B. (1991) A semiparametric test of fixed costs of labor adjustment, Preprint.
Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematik*, 31, 377–403.
Cristóbal Cristóbal, J. A., Faraldo Roca, P. and González Manteiga, W. (1987) A class of linear regression parameter estimators constructed by nonparametric estimation, *Annals of Statistics* 15, 603–609.
CURVDAT: STATCOM; Institut für Statistic Computing; Walter Köhler; Am Mühlrain 24 B D-6903; Neckargemünd.
Deheuvels, P. (1977) Estimation non paramétrique de la densité par histogrammes généralisés, *Review de Statistiquee Apliqueé*, 25, 5–42.
Deheuvels, P. and Hominal, P. (1980) Estimation automatiquee de la densité, *Review de Statistique Appliquee*, 28, 25–55.
Davidson and MacKinnon (1981) Several model specification tests in the presence of alternative hypotheses, *Econometrica* 49, 781–793.
Delgado, M. A. (1989a) Asymptotically efficient fully iterative nonlinear weighted least squares in the presence of heteroskedasticity of unknown form, Preprint.
—— (1989b) Semiparametric generalised least squares estimation in the multivariate nonlinear regression model, *Econometric Theory* (forthcoming).
—— (1990a) Bounded influence regression in the presence of heteroskedasticity of unknown form, *Nonparametric Functional Estimation and Related Topics* (G. Roussas ed.), Dordrecht: Kluwer Academic Publishers.
—— (1990b) Computing nonparametric functional estimates in semiparametric problems, Preprint.
Delgado, M.A. and Kniesner, T. (1990) Semiparametric versus parametric models for count data: modeling the causes of sickness spells. Preprint.
Delgado, M. A. and Stengos, T. (1990a) Semiparametric specification testing of non-nested econometric models, Preprint.
—— (1990b) N-Kernel: A review, *Journal of Applied Econometrics*, 5, 299–304.
Devroye, L. (1978) The uniform convergence of nearest neighbor regression function estimators and their application in optimization, *IEEE Transactions on Information Theory*, IT-24, 142–151.

—— (1982) Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression estimates, *Zeitschrift fur Wahrsheinlichkeitstheorie und verwandte Gebiete*, 61, 467–481.
—— (1987) *A Course in Density Estimation*. Boston: Birkhauser.
Devroye, L. and Györfi (1985) *Nonparametric Density Estimation: The L_1 View*. New York: J. Wiley and Sons.
Devroye, L. and Penrod, C. S. (1986) The strong uniform consistency of multivariate variable kernel estimates, *The Canadian Journal of Statistics*, 14, 211–219.
Devroye, L. and Wagner, T. J. (1980) Distribution-free consistency results for in nonparametric discrimination and regression function estimation, *Annals of Statistics*, 8, 231–239.
Diebold, F. X. and Nason, J. A. (1990) Nonparametric exchange rate prediction?, *Journal of International Economics*, 28, 315–332.
Doukhan, P. and Ghindès, M. (1980) Estimations dans le processus '$x_{n+1} = f(x_n) + \varepsilon_n$'. *C.R. Acad. Sci. Paris, Sér. A.*, 290, 921–923.
Duin, R. P. W. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Transactions on Computers*, C-25, 1175–1179.
Epanechnikov, V. A. (1969) Nonparametric estimation of a multivariate probability density, *Theory of Probability and its Applications*, 14, 153–158.
Engle, R. F., Granger, W. J., Rice, J. A. and Weiss, A. (1986) Semiparametric estimates of the relationship between weather and electricity sales, *Journal of the American Statistical Association*, 81, 310–320.
Eubank, R. and Spiegelman, S. (1990) Testing the goodness-of-fit of linear models via regression techniques, *Journal of the American Statistical Association*, 85, 387–397.
Faraldo Roca, P. and González Manteiga, W. (1985) On efficiency of a new class of linear regression estimates obtained by preliminary non-parametric regression, in *New Perspectives in Theoretical and Applied Statistics*, (M. Puri *et al.* eds) Wiley: New York.
Fix, E. and Hodges, J. L. (1951) Discriminatory analysis, nonparametric estimation: consistency properties, *Report Number 4, Project no. 21-49-004*, USAF School of Aviation Medicine, Randoph Field, Texas.
Freedman, D. and Diaconis, P. (1981a) On the maximum deviation between the histogram and the underlying density, *Zeitschrift fur Wahrsheinlichkeitstheorie und verwandte Gebiete*, 58, 139–157.
—— (1981b) On the histogram as a density estimator: L_2 theory, *Zeftschrift fur Wahrsheinlichkeitstheorie und verwandte Gebiete*, 58, 139–157.
Friedman, J. H., Baskett, F. and Shustek, L. J. (1975) An algorithm for finding nearest neighbors, *IEEE Transactions on Computers C-24*, 1149–1158.
Fryer, M. J. (1977) A review of some nonparametric methods of density estimation, *Journal of the Institute of Mathematics and its Applications*, 20, 335–354.
Fukunaga, K. (1972) *Introduction to Statistical Pattern Recognition*, New York: Academic Press.
GAUSS: *Aptech Systems Inc.*, Kent, WA.
Gasser, T. and Müller, H. G. (1979) Kernel estimation of regression functions, in *Smoothing Techniques for Curve Estimation*, (T. Gasser and M. Rosenblatt, eds) Lecture Notes in Mathematics 757, 23–68, Heidelberg: Springer-Verlag.
Gessaman, M. P. (1970) A consistent nonparametric multivariate density estimator based on statistically equivalent blocks, *Annals of Mathematical Statistics* 41, 1344–1346.
Gonzalez Manteiga, W. and Cadarso Suárez, C. M. (1990) Linear regression with randomly right-censored data using prior nonparametric estimation, *Nonparametric Functional Estimation and Related Topics* (G. Roussas ed.) Dordrecht: Kluwer Academic Publishers.
Green, P., Jennison, C. and Seheult, A. (1985) Analysis of field experiments by least squares smoothing, *Journal of the Royal Statistical Society, Series B*, 47, 299–315.

Györfi, L. (1987) Density estimation from dependent sample, in *Statistical Data Analysis based on the L_1 Norm and Related Methods*. (Y. Dodge ed.) Amsterdam: North-Holland.
—— (1991) Universal consistencies of a regression estimate for unbounded regression functions, in *Nonparametric Functional Estimation and Related Topics* (G. Roussas ed.) Dordrecht: Kluwer Academic Publishers.
Habbema, J. D. F., Hermans, J. and Remme, J. (1978) Variable kernel density estimation in discriminant analysis, *Compstat 1978, Proceedings in Computational Statistics*, Vienna: Physica Verlag.
Hall, P. (1983a) Large sample optimality of least squares cross-validation in density estimation, *Annals of Statistics*, 11, 1156–1174.
—— (1983b) Asymptotic theory of minimum integrated square error for multivariate density estimation, *Proceedings of the Sixth International Symposium on Multivariate Analysis*, Pittsburg.
Hall, P. and Hart, J. D. (1989a) Convergence rates in density estimation for data from infinite-order moving average processes, Preprint.
—— (1989b) Nonparametric estimation with long-range dependence, Preprint.
—— (1990) Bootstrap tests for the difference between means in nonparametric regression, *Journal of the American Statistical Association*, 85, 1039–1049.
Hand, D. J. (1982) *Kernel Discriminant Analysis*. Chichester: Research Studies Press.
Hannan, E. J. (1963) Regression for time series, in *Time Series Analysis* (ed. M. Rosenblatt) New York: Wiley.
—— (1965) The estimation of relationships involving distributed lags, *Econometrica*, 33, 206–224.
Hannan, E. J. and Terrell, R. D. (1973) Multiple equation systems with stationary errors, *Econometrica*, 41, 299–320.
Härdle, W. (1984) Robust regression function estimation, *Journal of Multivariate Analysis* 14, 169–180.
—— (1987) Resistant smoothing using the fast Fourier transform, Statistical Algorithm 222, *Applied Statistics* 36, 104–111.
—— (1990) *Applied Nonparametric Regression*, Cambridge: Cambridge University Press, Econometric Society Monographs.
Härdle, W. and Jerison, M. (1988) Evolution of Engle curves over time, Technical Report, University of Bonn.
Härdle, W. and Marron, J. S. (1985a) Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression, *Biometrika*, 72, 481–484.
—— (1985b) Optimal bandwidth selection in nonparametric function estimation, *Annals of Statistics*, 13, 1465–1481.
—— (1990) Comparing nonparametric versus parametric regression fits, Preprint.
Härdle, W. and Stoker, T. (1989) Investigating smooth multiple regression by the method of average derivatives, *Journal of the American Statistical Association*, 84, 986–995.
Härdle, W. and Tsybakov, A. B. (1990) Robust nonparametric regression with simultaneous scale curve estimation, *Annals of Statistics* 16, 120–135.
Härdle, W. and Vieu, P. (1989) Nonparametric prediction by the kernel method, Preprint.
Härdle, W., Hart, J., Marron, J. S. and Tsybakov, A. B. (1989) Bandwidth choice for average derivative estimation, Preprint.
Hart, J. D. and Vieu, P. (1990) Data driven bandwidth choice for density estimation based on dependent data, *Annals of Statistics*, 18, 873–890.
Hartigan, J. A. and Hartigan, P. M. (1985) The dip test of unimodality, *Annals of Statistics*, 13, 70–84.
Harvey, A. C. and Robinson, P. M. (1988) Efficient estimation of nonstationary time series regression, *Journal of Time Series Analysis*, 9, 201–214.

Hassani, S., Sarda, P. and Vieu, P. (1986) Approche non paramétrique en théorie de la fiabilité, *Revue de Statistique Appliqueés*, 35.
Heckman, N. E. (1986) Spline smoothing in a partly linear model, *Journal of the Royal Statistical Society*, Series B, 48, 244–248.
Hidalgo, F. J. (1989) Adaptive estimation in time series regression models with heteroskedasticity of unknown form, *Journal of Time Series Analysis*, (forthcoming).
—— (1990) Adaptive semiparametric estimation in the presence of autocorrelation of unknown form, *Econometric Theory*, (forthcoming).
Hildenbrand, K. and Hildenbrand, W. (1980) On the mean income effect: a data analysis of the U.K. family expenditure family, in *Contributions to Mathematical Economics*, W. Hildenbrand and A. Mas-Colell (eds) New York: New-Holland.
Hill, J. D. (1969) A search technique for multimodal surfaces, *IEEE Transactions on Systems, Science and Cybernetics*, SSC-5, 2–8.
Hsieh, D. and Manski, C. (1987) Monte Carlo evidence on adaptive maximum likelihood estimation, *Annals of Statistics*, 15, 541–551.
Izenman, A. J. (1991) Recent developments in nonparametric density estimation, *Journal of the American Statistical Association*, 86, 205–224.
International Mathematical and Statistical Libraries, Inc. (1984) *IMSL Library: FORTRAN Subroutines for Mathematics and Statistics* (ed. 9.2).
—— (1987) *STAT|LIBRARY* (Version 1.0).
Jarvis, R. A. (1970) Adaptive global search in a time-variant environment using a probabilistic automaton with pattern recognition supervision, *IEEE Transactions on Systems, Science and Cybernetics*, SSC-6, 209–216.
Johnston, G. J. (1982) Probabilities of maximal deviations for nonparametric regression function estimates, *Journal of Multivariate Analysis*, 12, 402–414.
King, E. C. (1989) *A Test for the Equality of Two Regression Curves*, Ph.D. Thesis, Dep. of Stat., Texas A & M.
Kogure, A. (1987) Asymptotically optimal cells for a histogram, *Annals of Statistics*, 15, 1023–1030.
Kreiss, J. P. (1987) On adaptive estimation of stationary ARMA processes, *Annals of Statistics*, 15, 112–133.
Lecoutre, J. P. (1986) The histogram with random partition, in *New Perspectives in Theoretical and Applied Statistics*, (M. Puri *et al.* eds) Wiley: New York.
Lee, B.-J. (1991) A nonparametric specification test using a kernel estimation method, Preprint.
Lee, D. K. C. (1989) Semiparametric analysis of the 'surprise' consumption function, Preprint.
—— (1990a) Elasticity, fuel efficiency and attribute demand: a semiparametric hedonic approach, Preprint.
—— (1990b) Consumption, growth, interest rates, inflation and ARCH effect of an unknown form, Preprint.
—— (1990c) Cross-validation in semiparametric models: some Montecarlo results, *Journal of Statistical Simulation and Computation*, 37, 171–187.
—— (1992) N-Kernel and XploRe, *Journal of Economic Surveys*, 6, 89–105.
Lee, L. F. (1990) Efficient semiparametric scoring estimation of sample selection models, Preprint.
—— (1991) Semiparametric nonlinear least squares estimation of truncated regression models, *Econometric Theory*, (forthcoming).
Leonard, T. (1978) Density estimation, stochastic processes, and prior information, (with discussion), *Journal of the Royal Statistical Society, Series B*, 40, 113–146.
Loftsgaarden, D. O. and Quesenberry, C. P. (1965) A nonparametric estimate of a multivariate density function, *Annals of Mathematical Statistics*, 36, 1049–1051.

Li, K. C. (1984) Consistency of nearest neighbor estimates in non-parametric regression, *Annals of Statistics* 12, 230–240.
—— (1985) From Stein's unbiased risk estimates to the method of generalized cross-validation, *Annals of Statistics*, 13, 1352–1377.
—— (1987) Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set, *Annals of Statistics*, 15, 958–975.
Mack, Y. P. (1981) Local properties of *k-NN* regression estimates, *SIAM Journal of Algebraic Discrete Methods*, 2, 311–323.
Mack, Y. P. and Rosenblatt, M. (1979) Multivariate k-nearest neighbor density estimates, *Journal of Multivariate Analysis*, 9, 1–15.
Manski, C. F. (1984) Adaptive estimation of non-linear regression models, *Econometric Reviews* 3, 145–194.
MAT LAB: *The MATH WORKS Inc.*, 21 Eliot Street, South Natick, MA 01760.
McMurtry, G. J. and Fu, K. S. (1966) A variable structure automaton used as a multi modal searching technique, *IEEE Transactions in Automatic Control*, AC-11, 379–387.
McQueen, J. B. (1990) *N-Kernel*, Non-standard Statistical Software, Santa Monica.
Moore, D. S. and Yackel, J. W. (1977) Consistency properties of nearest neighbor density function estimates, *Annals of Statistics*, 5, 143–154.
Müller, H. G. and Stadtmüller, U. (1987) Estimation of heteroskedasticity in regression analysis, *Annals of Statistics*, 15, 610–625.
Nadaraya, E. A. (1964) On estimating regression, *Theory of Probability and its Applications* 9, 141–142.
Ng, P. T. and Sickles, R. C. (1990) 'XploRe'-ing the world of nonparametric analysis, *Journal of Applied Econometrics* 5, 293–298.
Newey, W. K. (1989) Locally efficient, residual-based estimation of nonlinear simultaneous equations, Preprint.
—— (1990a) Efficient instrumental variable estimation of nonlinear models, *Econometrica* 58, 809–837.
—— (1990b) Semiparametric efficiency bounds, *Journal of Applied Econometrics*, (forthcoming).
—— (1990c) Efficient estimation of semiparametric models via moment restrictions, Preprint.
—— (1991) Series estimators of regression functionals, Preprint.
Newey, W. K. and Powell, J. L. (1987a) Efficient estimation of type I censored regression models under conditional quantile and symmetry restrictions, Preprint.
—— (1987b) Efficient estimation of Tobit models under conditional quantile and symmetry restrictions, Preprint.
Newey, W. K., Powell, J. L. and Walker, J. R. (1990) Semiparametric estimation of selection models, *American Economic Review, Papers and Proceedings*, 80, 324–328.
Newton, H. J. (1988) *TIMESLAB: A Time Series Analysis Laboratory*, Belmont: Wadsworth.
Pagan, A. R. and Ullah, A. (1988) The econometric analysis of models with risk terms, *Journal of Applied Econometrics* 3, 87–105.
Parzen, E. (1962) On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, 33, 1065–1076.
Powell, J. L. (1989) Semiparametric estimation of censored regression models, Preprint.
Powell J. L., Stock, J. H. and Stoker, T. M. (1989) Semiparametric estimation of index coefficients, *Econometrica* 57, 1403–1430.
Prakasa Rao, B. L. S. (1983) *Nonparametric Functional Estimation*, Orlando: Academic Press.
Prescott, D. M. and Stengos, T. (1988) Do asset markets overlook exploitable nonlinearities? The case of gold, Preprint.

Priestley, M. B. and Chao, M. T. (1972) Nonparametric function fitting, *Journal of the Royal Statistical Society, Series B*, 34, 385–392.
Rèvèsz, P. (1972) On empirical density function, *Periodica Mathematica Hungarica*, 2, 85–110.
Rice, J. A. (1984) Bandwidth choice for nonparametric regression, *Annals of Statistics*, 12, 1215–1230.
—— (1986) Convergence rates for partially splined models, *Statistics and Probability Letters*, 4, 203–208.
Robinson, P. M. (1976) The estimation of linear differential equations with constant coefficients, *Econometrica*, 44, 751–764.
—— (1983) Nonparametric estimators for time series, *Journal of Time Series Analysis*, 4, 185–207.
—— (1984) Robust nonparametric autoregression, *Lecture Notes in Statistics* 26, 247–255.
—— (1986) Nonparametric estimation of time-varying parameters, in *Analysis and Forecasting of Economic Structural Change*, Amsterdam: North-Holland.
—— (1987a) Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form, *Econometrica* 55, 531–548.
—— (1987b) Time series residuals with application to probability density estimation, *Journal of Time Series Analysis*, 8, 329–344.
—— (1987c) Adaptive estimation of heteroskedastic regression models, *Revista de Econometria*, 7, 5–28.
—— (1987d) Nonparametric function estimates for long-memory time series in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, (W. Barnett *et al.* eds) New York: CUP.
—— (1988a) Root-n-consistent semiparametric regression, *Econometrica* 56, 931–954.
—— (1988b) Semiparametric econometrics: a survey, *Journal of Applied Econometrics* 3, 35–51.
—— (1989) Hypothesis testing in semiparametric and nonparametric models for econometric time series, *Review of Economic Studies* 56, 511–534.
—— (1990a) Best nonlinear three-stage least squares of certain econometric models, *Econometrica*, (forthcoming).
—— (1990b) Automatic frequency-domain inference on semiparametric and nonparametric models, *Econometrica*, (forthcoming).
—— (1991) Consistent nonparametric entropy-based testing, *Review of Economic Studies*, (forthcoming).
Rose, R. L. (1978) *Nonparametric Estimation of Weights in Least-Squares Regression Analysis*, Thesis, University of California at Davis.
Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *Annals of Mathematical Statistics*, 27, 832–837.
—— (1969) Conditional probability density and regression estimators, in *Multivariate Analysis II*, (ed. P. R. Krishnaiah), New York: Academic Press, 25–31.
—— (1971) Curve estimates, *Annals of Statistics*, 42, 1815–1842.
—— (1979) Global measures of deviation for kernels and nearest neighbor density estimates, in *Smoothing Techniques for Curve Estimation*, (eds T. Gasser and M. Rosenblatt) *Lecture Notes in Mathematics 757*, 181–190, Berlin: Springer-Verlag.
Roussas, G. G. (1969) Nonparametric estimation of the transition distribution of a Markov process, *Annals of Mathematical Statistics*, 40, 1386–1400.
—— (1988) Nonparametric estimation in mixing sequences of random variables, *Journal of Statistical Planning and Inference*, 18, 135–149.
Royall, R. M. (1966) *A class of nonparametric estimators of a smooth regression function*, Thesis, Stanford University
Rudemo, M. (1982) Empirical choice of histogram and kernel density estimators, *Scandinavian Journal of Statistics*, 9, 65–78.

Samarov, A. M. (1990) Exploring regression structure using nonparametric functional estimation, Preprint.
Schick, A. (1986) On asymptotically efficient estimation in semiparametric models, *Annals of Statistics*, 14, 1139–1151.
Schuster, E. F. (1972) Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, *Annals of Mathematical Statistics*, 43, 84–88.
Schuster, E. F. and Gregory, C. G. (1981) On the nonconsistency of maximum likelihood nonparametric density estimators, in *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (ed. W. F Eddy) New York: Springer Verlag.
Scott, D. W. (1979) On optimal and data based histograms, *Biometrika*, 66, 605–610.
—— (1985a) Average shifted histograms: effective nonparametric density estimators in several dimensions, *Annals of Statistics*, 13, 1024–1040.
—— (1985b) Frequency polygons: theory and applications, *Journal of the American Statistical Association*, 80, 348–354.
Scott, D. W., Tapia, R. A. and Thompson, J. R. (1977) Kernel density estimation revisited, *Nonlinear Analysis*, 1, 339–372.
Singpurwalla, N. D. and Wong, Y. (1983) Estimation of the failure rate: a survey of nonparametric methods. Part I: non Bayesian methods, in *Communications in Statistical Theory and Mathematics*, 12, 559–588 .
Silverman, B. W. (1981) Using kernel density estimates to investigate multimodality, *Journal of the Royal Statistical Society, Series B*, 43, 97–99.
—— (1982) Kernel density estimation using the fast Fourier transform, Statistical Algorithm 175, *Applied Statistics* 31, 93–97.
—— (1983) Some properties of a test for multimodality based on kernel density estimates, in *Probability, Statistics and Analysis*, (eds J. F. C. Kingman and G. E. H. Reuter), Cambridge: Cambridge University Press, 248–259.
—— (1986) *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.
Silveira, G. (1990) L_1-strong consistency for density estimates in dependent samples, in *Nonparametric Functional Estimation and Related Topics* (G. Roussas ed.), Dordrecht: Kluwer Academic Publishers.
Speckman, P. (1988) Kernel smoothing in partially linear models, *Journal of the Royal Statistical Society, Series B*, 50, 413–446.
Spiegelman, C. H. (1976) *Two techniques for estimating treatment effects in the presence of hidden variables: adaptive regression and a solution to Riersol's problem*. Thesis, Northwestern University.
Spiegelman, C. H. and Sacks, J. (1980) Consistent window estimation in nonparametric regression, *Annals of Statistics*, 8, 240–246.
Steigerwald, D. (1990) Adaptive estimation in time series models, Preprint.
Stein, C. (1956) Efficient nonparametric testing and estimation, in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley: University of California Press.
Stock, J. H. (1989) Nonparametric policy analysis, *Journal of the American Statistical Association*, 84, 567–577.
—— (1990) Nonparametric policy analysis: an application to estimating hazardous waste cleanup benefits in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, (W. Barnett *et al.* eds) New York: CUP.
Stoker, T. M. (1989) Tests of additive derivative constraints, *Review of Economic Studies*, 56, 535–552.
Stone, C. J. (1975) Adaptive maximum likelihood estimation of a location parameter, *Annals of Statistics* 3, 267–284.
—— (1977) Consistent nonparametric regression (with discussion), *Annals of Statistics* 5, 595–645.

—— (1980) Optimal rates of convergence for nonparametric estimators, *Annals of Statistics*, 8, 1348–1360.
—— (1982) Optimal rates of convergence for nonparametric regression, *Annals of Statistics*, 10, 1040–1053.
—— (1984) An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics*, 12, 1285–1297.
Stute, W. (1984) Asymptotic normality of nearest neighbor regression function estimates, *Annals of Statistics*, 12, 917–926.
Tapia, R. A. and Thompson, J. R. (1978) *Nonparametric Probability Density Estimation*, Baltimore, MD: John Hopkins University Press.
Tarter, M. E. and Kronmal, R. A. (1976) An introduction to the implementation and theory of nonparametric density estimation, *American Statistician*, 30, 105–112.
Tran, L. T. (1989) The L_1 convergence of kernel density estimates under dependence, *Canadian Journal of Statistics*, 17, 197–208.
Tsybakov, A. B. (1982) Robust estimates of a function, *Problems, Information and Transmission* 18, 190–201.
Van Ryzin, J. (1973) A histogram method of density estimation, *Communications in Statistics*, 2, 493–506.
Watson, G. S. (1964) Smooth regression analysis, *Sankhya* A 26, 359–372.
Wegman, E. J. (1982). Density estimation, in *Encyclopedia of Statistical Sciences, Vol 2* (eds S. Kotz and N. L. Johnson) 309–315, New York: Wiley.
Wertz, W. and Schneider, B. (1979) Statistical density estimation: a bibliography, *International Statistical Review*, 47, 155–175.
Whistler, D. (1989) Semi-parametric ARCH estimation of intra-daily exchange volatility, Preprint.
Whittle, P. (1958) On smoothing of probability densities, *Journal of the Royal Statistical Society*, 20, 334–343.
Woodroofe, M. (1970) On choosing a delta sequence, *Annals of Mathematical Statistics*, 41, 1665–1671.
Wooldridge, J. (1990) A test for functional form against nonparametric alternatives, Preprint.
Yakowitz, S. (1985) Nonparametric density estimation, prediction and regression for Markov's sequences, *Journal of the American Statistical Association*, 80, 215–221.
—— (1987) Nearest neighbor methods for time series analysis, *Journal of Time Series Analysis*, 8, 235–247.
Yatchew, A. (1990) Nonparametric regression tests based on least squares, Preprint.
Yatchew, A. and Bos, L. (1991) Nonparametric regression model tests, Preprint.
Yang, S. (1981) Linear functions of concomitants of order statistics with application to nonparametric estimation of a regression function, *Journal of the American Statistical Association*, 76, 658–662.